MedIR @ SIGIR 2014

July 11, 2014

Held in Gold Coast, Australia

Medical Information Retrieval Workshop at SIGIR 2014

Editors:
Lorraine Goeuriot
Gareth J.F. Jones
Liadh Kelly
Henning Müller
Justin Zobel

Preface

Medical information is accessible from diverse sources including the general web, social media, journal articles, and hospital records; users include patients and their families, researchers, practitioners and clinicians. Challenges in medical information retrieval include: diversity of users and user ability; variations in the format, reliability, and quality of biomedical and medical information; the multimedia nature of data; and the need for accuracy and reliability. The objective of the Medical Information Retrieval workshop is to provide a forum to enable the progression of research in medical information retrieval to provide enhanced search services for all users with interests in medical information search. The workshop aims to bring together researchers interested in medical information search with the goal of identifying specific research challenges that need to be addressed to advance the state-of-the-art and to foster interdisciplinary collaborations towards the meeting of these challenges. To enable this, we encouraged participation from researchers in all fields related to medical information search including mainstream information retrieval, but also natural language processing, multilingual text processing, and medical image analysis.

The organizers would like to thank Dr Karin Verspoor (University of Melbourne, Australia) for giving a keynote talk at the workshop, paper authors for their invaluable contribution, the members of the program committee for their help in the reviewing process, and SIGIR for hosting the event.

Lorraine Goeuriot
Gareth J.F. Jones
Liadh Kelly
Henning Müller
Justin Zobel

July 2014

Organization

Program Committee Chairs

Lorraine Goeuriot, Dublin City University, Ireland
Gareth JF Jones, Dublin City University, Ireland
Liadh Kelly, Dublin City University, Ireland
Henning Müller, University of Applied Sciences Western Switzerland
Justin Zobel, University of Melbourne, Australia

Program Committee

Eiji Aramaki, Kyoto University, Japan
Celia Boyer, Health on the Net, Switzerland
Ben Carterette, University of Delaware, USA
Allan Hanbury, Vienna University of Technology, Austria
William Hersh, Oregon Health and Science University, USA
Jung-Jae Kim, Nanyang Technological University, Singapore
Gang Luo, University of Utah, USA
Iadh Ounis, University of Glasgow, UK
Patrick Ruch, HES-SO, Switzerland
Stefan Schulz, Medical University Graz, Austria
Karin Verspoor, NICTA, Australia
Ellen Voorhees, NIST, USA
Ryen White, Microsoft Research, USA
Elad Yom-Tov, Microsoft Research, USA
Pierre Zweigenbaum, LIMSI, France

Table of Contents

Preface III

Organization IV

Table of Contents V

Overview

Report on the SIGIR 2014 Workshop on Medical Information Retrieval (MedIR) 1

Keynote

Practice-based Evidence in Medicine: Where Information Retrieval Meets Data Mining 4
Karin Verspoor

Papers

Designing for Health Exploratory Seeking Behaviour 5
Patrick Cheong-Iao Pang, Karin Verspoor, Shanton Chang and Jon Pearce.

Evaluation of Coreference Resolution for Biomedical Tex 9
Miji Choi, Karin Verspoor and Justin Zobel.

Retrieving Attitudes: Sentiment Analysis from Clinical Narratives 12
Yihan Deng, Matthaeus Stoehr and Kerstin Denecke.

Why Assessing Relevance in Medical IR is Demanding 16
Bevan Koopman and Guido Zuccon.

Multi-modal relevance feedback for medical image retrieval 20
Dimitrios Markonis, Roger Schaer and Henning Müller.

A Joint Local-Global Approach for Medical Terminology Assignment 24
Liqiang Nie, Mohammad Akbari, Tao Li and Tat-Seng Chua.

Exploring Clustering Based Knowledge Discovery towards Improved Medical Diagnosis 28
Rajendra Prasath and Philip O'Reilly.

Integrating Understandability in the Evaluation of Consumer Health Search Engines 32
Guido Zuccon and Bevan Koopman.

Report on the SIGIR 2014 Workshop on Medical Information Retrieval (MedIR)

Lorraine Goeuriot
Dublin City University
Ireland
lorraine.goeuriot@imag.fr

Liadh Kelly
Dublin City University
Ireland
liadh.kelly@scss.tcd.ie

Gareth J.F. Jones
Dublin City University
Ireland
gjones@computing.dcu.ie

Henning Müller
HES-SO Valais
Switzerland
henning.mueller@hevs.ch

Justin Zobel
University of Melbourne
Australia
jzobel@unimelb.edu.au

ABSTRACT

The workshop on Medical Information Retrieval took place at SIGIR 2014 in Gold Coast, Australia on July 11. The workshop included eight oral presentations of referred papers and an invited keynote presentation. This allowed time for lively discussions among the participants. These showed the significant interest in the medical information retrieval domain and the many research challenges arising in this space which need to be addressed to give added value to the wide variety of users that can profit from medical information search, such as patients, general health professionals and specialist groups such as radiologists who mainly search for images and image related information.

1. INTRODUCTION

Medical information retrieval refers to methodologies and technologies that seek to improve access to medical information archives via a process of information retrieval (IR). Such information is now potentially accessible from many sources including the general web, social media, journal articles, and hospital records. Health-related content is one of the most searched-for topics on the Internet, and as such this is an important domain for research in information retrieval.

Medical information is of interest to a wide variety of users, including patients and their families, researchers, general practitioners and clinicians, and practitioners with specific expertise such as radiologists. There are several dedicated services that seek to make this information more easily accessible, such as Health on the Net's medical search systems for the general public and medical practitioners: http://www.hon.ch/. Despite the popularity of the medical domain for users of search engines, and current interest in this topic within the IR research community, development of search and access technologies remains particularly challenging and under explored.

One of the central issues in medical information search is the diversity of the users of these services with corresponding differences in types and scopes of their individual needs. Their information needs will be associated with varied categories and purposes, they will typically have widely varying levels of medical knowledge, and, important in some settings, they will have differing language skills.

These challenges can be summarized as follows:

1. Varying information needs: While a patient with a recently diagnosed condition will generally benefit most from simple or introductory information on the disease and its treatment, a patient living with or managing a condition over a longer term will generally be looking for more advanced information, or perhaps support groups and forums. In a similar way, a general practitioner might require basic information quickly while advising a patient, but more detailed information if deciding on a course of treatment, while a specialist clinician might look for an exhaustive list of similar cases or research papers relating to the condition of a patient that they are currently seeking to advise. Understanding various types of users and their information needs is one of the cornerstones of medical information search, while adapting IR to best address these needs to develop effective, potentially personalized systems is one of its greatest challenges.

2. Varying medical knowledge: The different categories of users of medical information search systems will have widely varying levels of medical knowledge, and indeed the medical knowledge of different individuals within a user category can also vary greatly. This affects the way in which individuals pose search queries to systems and also the level of complexity of information which should be returned to them or the type of support in understanding / disambiguating returned material which will be required.

3. Varying language skills: Given that much medical content is written only in the English language, research to date in medical information search has predominantly focused on monolingual English retrieval. However, given the large number of non-English speakers on the Internet and the lack of content in their native lan-

MedIR 2014, July 11, 2014, Gold Coast, Australia.

guage, effective support for them to search English language sources is highly desirable. The Internet in particular has affected the patient-physician relationship, and providing relevant, reliable information to patients in their own language is a key to alleviate such challenging situations and reduce instances of phenomenon such as cyberchondria.

In addition, the format, reliability, and quality of biomedical and medical information varies greatly. A single health record can contain clinical notes, technical pathology data, images, and patient-contributed histories, and may be linked by a physician to research papers. The importance of health and medical topics and their impact on people's everyday lives makes the need for retrieval of accurate and reliable information especially important. Determining the likely reliability of available information is challenging. Finally, as with IR in general, the evaluation of medical search tools is vital and challenging. For example, there are no established or standardized baselines or evaluation metrics, and limited availability of test collections. Further discussion and progression on this topic would be beneficial to the community.

2. THEME AND PURPOSE OF THE WORKSHOP

The objective of the workshop was to provide a forum to enable the progression of research in medical IR seeking to provide enhanced search services for all users with interests in medical information search. The workshop aimed to bring together researchers interested in medical information search with the goal of identifying specific research challenges that need to be addressed to advance the state-of-the-art and to foster interdisciplinary collaborations towards the meeting of these challenges. To enable this, we encouraged participation from researchers in all fields related to medical information search including mainstream IR, but also natural language processing, multilingual text processing, and medical image analysis.

Topics of interest included but are were not limited to:

- Users and information needs
- Semantics and natural language processing (NLP) for medical IR
- Reliability and trust in medical IR
- Personalised search
- Evaluation of medical IR
- Multilingual issues in medical IR
- Multimedia technologies in medical IR
- The role of social media in medical IR

3. KEYNOTE - DR KARIN VERSPOOR

The keynote talk was given by Dr Karin Verspoor (University of Melbourne, Australia), on "Practice-based Evidence in Medicine: Where Information Retrieval Meets Data Mining" [7]. A new approach in medical practice is emerging thanks to the increasing availability of large-scale clinical data in electronic form. In practice-based evidence, the clinical record is mined to identify patterns of health characteristics, such as diseases that co-occur, side-effects of treatments, or more subtle combinations of patient attributes that might explain a particular health outcome. This approach contrasts with what has been the standard of care in medicine, evidence-based practice, in which treatment decisions are based on (quantitative) evidence derived from targeted research studies, specifically, randomised controlled trials. Advantages of consulting the clinical record for evidence rather than relying solely on structured research include avoiding the selection bias of the inclusion criteria for a clinical trial and monitoring of longer-term outcomes and effects. The two approaches are, of course, complementary - a hypothesis derived from large-scale data mining could in turn form the starting point for the design of a clinical trial to rigorously investigate that hypothesis. Information retrieval can play an important role in both approaches to collecting medical evidence. However, the use of information retrieval methods in collecting practice-based evidence requires moving away from traditional document-oriented retrieval as the end goal in itself, to viewing that retrieval as an intermediate step towards knowledge discovery and population-scale data mining. Furthermore, it may require the development of more context-specific retrieval strategies, designed to identify specific characteristics of interest and support particular tasks in the medical context.

4. PRESENTED PAPERS

Of the twenty papers submitted to the workshop, eight were selected for inclusion in the workshop proceedings and for presentation at the workshop:

- Patrick Cheong-Iao Pang, Karin Verspoor, Shanton Chang and Jon Pearce. Designing for Health Exploratory Seeking Behaviour [5]
- Miji Choi, Karin Verspoor and Justin Zobel. Evaluation of Coreference Resolution for Biomedical Text [1]
- Yihan Deng, Matthaeus Stoehr and Kerstin Denecke. Retrieving Attitudes: Sentiment Analysis from Clinical Narratives [2]
- Bevan Koopman and Guido Zuccon. Why Assessing Relevance in Medical IR is Demanding []
- Dimitrios Markonis, Roger Schaer and Henning MÃijller. Multi-modal relevance feedback for medical image retrieval [3]
- Liqiang Nie, Mohammad Akbari, Tao Li and Tat-Seng Chua. A Joint Local-Global Approach for Medical Terminology Assignment [4]
- Rajendra Prasath and Philip O'Reilly. Exploring Clustering Based Knowledge Discovery towards Improved Medical Diagnosis [6]
- Guido Zuccon and Bevan Koopman. Integrating Understandability in the Evaluation of Consumer Health Search Engines [9]

5. DISCUSSION SESSION

The discussion sessions started with a brainstorming activity to identify the key challenges in medical IR. The two main areas identified were the lack of available data sets and the need for better evaluation. Two groups were formed to discuss these two topics.

The first group discussed the lack of data sets. One of the reasons for this is the limited amount of publicly available data (i.e. clinical data, query logs, etc.). Aside from the patient related issues of confidentiality and privacy, medical data being very varied and changing, getting representative and up-to-date data sets is very challenging. These variations can be found at different levels. The level of specialization and targeted readers is the first one: consumer information varies greatly from clinical practice information. Then, linguistic variations such as shifting vocabulary are impacting information extraction (IE) and IR results. In order to deal with these changing characteristics, what could the value of abstraction into controlled vocabularies be? Moreover, controlled vocabulary would help in alleviating ambiguity. But how can it be efficiently incorporated into a retrieval approach? Concept-based representation of data and indexing are investigated but their efficiency in IR is still to be proven. Finally, some modalities are very specific to the medical domain, such as temporality, negativity, and patients' characteristics in clinical data such as age, gender, co-morbidities, etc. To understand and automatically process these, training data is necessary (raw data and gold standard annotations), but is difficult and expensive to obtain.

The second group focused on the evaluation of medical information retrieval. They identified as the main issues the lack of evaluation campaigns and benchmarks for medical IR, and the lack of information on the few existing campaigns. Based on the experience of the group members, a few key challenges were focused on, in order to get more benchmarks, and improve their quality. Firstly, it is crucial to design realistic tasks, which involve a deep understanding of the users and their needs. Only once that has been done can the dataset be built, with realistic data. Along with the task and use case scenario, the evaluation scheme and the definition of relevance needs to be very carefully planned, in order to maximize the outcome of the task. In [2], relevance is modelled according to several relevance dimension factors: understandability, topicality, novelty, scope and reliability. For instance, a task focusing on IR for patients or laypeople would define relevance as mainly based on the topicality, the reliability and the understandability. These factors need to be taken into account during the relevance judgement and results evaluation stages [8]. This would allow personalization of the search, characterizing the users and their information needs. Lastly, benchmark creators should be incited to make their datasets available to the public, for specific tasks or any related research work.

6. CONCLUSIONS

This was the first SIGIR workshop on 'Medical Information Retrieval', and followed on nicely from the SIGIR 2013 medical workshop on 'Health Search and Discovery: Helping Users and Advancing Medicine'. The volume of interest in the workshop, both through the number of paper submissions and large number of workshop participants, highlight both the activity and interest in the medical information retrieval space within the community. The workshop provided greater insights into the active areas of research within this space and helped in progression of the many challenges facing the space. Special attention was paid to evaluation within this space and possibilities for progression within the data set creation and benchmarking initiatives discussed.

7. ACKNOWLEDGEMENTS

We would like to thank SIGIR 2014 for hosting the workshop. Thanks also go to the program committee (Eiji Aramaki, Kyoto University, Japan; Celia Boyer, Health on the Net, Switzerland; Ben Carterette, University of Delaware, USA; Allan Hanbury, Vienna University of Technology, Austria; William Hersh, Oregon Health and Science University, USA; Jung-Jae Kim, Nanyang Technological University, Singapore; Gang Luo, University of Utah, USA; Iadh Ounis, University of Glasgow, UK; Patrick Ruch, HES-SO, Switzerland; Stefan Schulz, Medical University Graz, Austria; Karin Verspoor, NICTA, Australia; Ellen Voorhees, NIST, USA; Ryen White, Microsoft Research, USA; Elad Yom-Tov, Microsoft Research, USA), paper authors and workshop attendees, without whom the workshop would not have been the success it was.

8. REFERENCES

[1] M. Choi, K. Verspoor, and J. Zobel. Evaluation of coreference resolution for biomedical text. In *Proceedings of the SIGIR workshop on Medical Information Retrieval (MEDIR 2014)*, 2014.

[2] Y. Deng, M. Stoehr, and K. Denecke. Retrieving attitudes: Sentiment analysis from clinical narratives. In *Proceedings of the SIGIR workshop on Medical Information Retrieval (MEDIR 2014)*, 2014.

[3] B. Koopman and G. Zuccon. Why assessing relevance in medical ir is demanding. In *Proceedings of the SIGIR workshop on Medical Information Retrieval (MEDIR 2014)*, 2014.

[4] L. Nie, M. Akbari, T. Li, and T.-S. Chua. A joint local-global approach for medical terminology assignment. In *Proceedings of the SIGIR workshop on Medical Information Retrieval (MEDIR 2014)*, 2014.

[5] P. C.-I. Pang, K. Verspoor, S. Chang, and J. Pearce. Designing for health exploratory seeking behaviour. In *Proceedings of the SIGIR workshop on Medical Information Retrieval (MEDIR 2014)*, 2014.

[6] R. Prasath and P. O'Reilly. Exploring clustering based knowledge discovery towards improved medical diagnosis. In *Proceedings of the SIGIR workshop on Medical Information Retrieval (MEDIR 2014)*, 2014.

[7] K. Verspoor. Practice-based evidence in medicine: Where information retrieval meets data mining. In *Proceedings of the SIGIR workshop on Medical Information Retrieval (MEDIR 2014), keynote*, 2014.

[8] Y. Zhang, J. Zhang, M. Lease, and J. Gwizdka. Multidimensional relevance modeling via psychometrics and crowdsourcing. In *Proceedings of SIGIR 2014*, 2014.

[9] G. Zuccon and B. Koopman. Integrating understandability in the evaluation of consumer health search engines. In *Proceedings of the SIGIR workshop on Medical Information Retrieval (MEDIR 2014)*, 2014.

Practice-based Evidence in Medicine: Where Information Retrieval Meets Data Mining

Karin M. Verspoor[1,2]
[1]Department of Computing and Information Systems
[2]Health and Biomedical Informatics Centre
The University of Melbourne
Melbourne, Victoria, Australia
karin.verspoor@unimelb.edu.au

1. INTRODUCTION

A new approach in medical practice is emerging thanks to the increasing availability of large-scale clinical data in electronic form. In *practice-based evidence* [5, 6], the clinical record is mined to identify patterns of health characteristics, such as diseases that co-occur, side-effects of treatments, or more subtle combinations of patient attributes that might explain a particular health outcome. This approach contrasts with what has been the standard of care in medicine, *evidence-based practice*, in which treatment decisions are based on (quantitative) evidence derived from targeted research studies, specifically, randomised controlled trials. Advantages of consulting the clinical record for evidence rather than relying solely on structured research include avoiding the selection bias of the inclusion criteria for a clinical trial and monitoring of longer-term outcomes and effects [5]. The two approaches are, of course, complementary — a hypothesis derived from large-scale data mining could in turn form the starting point for the design of a clinical trial to rigorously investigate that hypothesis.

Information retrieval can play an important role in both approaches to collecting medical evidence. However, the use of information retrieval methods in collecting practice-based evidence requires moving away from traditional document-oriented retrieval as the end goal in itself, to viewing that retrieval as an intermediate step towards knowledge discovery and population-scale data mining. Furthermore, it may require the development of more context-specific retrieval strategies, designed to identify specific characteristics of interest and support particular tasks in the medical context.

2. IR AND EVIDENCE-BASED PRACTICE

In evidence-based medicine, collection and meta-analysis of the published literature of clinical trials form the foundation of *systematic reviews* (e.g., Cochrane Reviews [1]). The production of such reviews has traditionally been done using painstaking exhaustive searches of the literature and human synthesis of published experimental results. It has been argued that automation is both necessary and possible [2, 7]. There is a clear role for information retrieval in this process, to identify publications relevant to a given review, although further structuring of the information within the documents retrieved is also needed [3].

A number of targeted search engines for the published biomedical literature have been developed that aim to improve search effectiveness for biomedical researchers [4]. Several incorporate the results of information extraction, such as named entity recognition for specific relevant entity types (e.g., drugs and diseases), with the objective of enabling concept-based indexing of the literature.

3. IR AND PRACTICE-BASED EVIDENCE

Data mining of electronic health records for medical evidence demands processing of the wealth of clinical data now recorded in natural language text. Transformation of this unstructured data into a structured representation is needed for incorporation of the information it contains into broader data mining. Many transformations can be cast as information retrieval tasks: for instance, identifying patients satisfying particular profiles (e.g., for recruitment into clinical trials or registries), or retrieval of case histories corresponding to specific treatment protocols. Development of general approaches to such tasks will likely require a mix of information retrieval and domain-specific information extraction.

4. CONCLUSION

The boundaries between information retrieval, information extraction, and data mining are blurring; bringing them together, in an activity commonly referred to as *text mining*, can result in heterogeneous methods that will enable sifting through the entirety of the clinical record, including both its unstructured and structured components. This in turn will enable clinical decision making based on data derived from large populations in the "laboratory" of the natural world.

5. REFERENCES

[1] Cochrane Collaboration. `http://www.cochrane.org`.
[2] T. Guy et al. The automation of systematic reviews. *BMJ*, 346, 2013.
[3] S. Kim et al. Automatic classification of sentences to support evidence based medicine. *BMC Bioinformatics*, 12(Suppl 2):S5, 2011.
[4] Z. Lu. Pubmed and beyond: a survey of web tools for searching biomedical literature. *Database*, baq036, 2011.
[5] T. Pincus and T. Sokka. Evidence-based practice and practice-based evidence. *Nat Clin Pract Rheum*, 2(3):114–115, 2006.
[6] N. H. Shah. Mining the ultimate phenome repository. *Nat Biotech*, 31(12):1095–1097, 2013.
[7] I. Shemilt et al. Pinpointing needles in giant haystacks: Use of text mining to reduce impractical screening workload in extremely large scoping reviews. *Research Synthesis Methods*, 2013. online preprint.

MedIR 2014 July 11, 2014, Gold Coast, Australia

Designing for Health Exploratory Seeking Behaviour

Patrick Cheong-Iao Pang[1,2], Karin Verspoor[1,3], Shanton Chang[1], Jon Pearce[1]
[1] Department of Computing and Information Systems, The University of Melbourne
[2] NICTA Victoria
[3] Health and Biomedical Informatics Centre, The University of Melbourne
Parkville, Victoria 3010, Australia
cipang, karin.verspoor, shanton.chang, j.pearce@unimelb.edu.au

ABSTRACT
The Internet has become a popular source of health information. However, in-depth understanding of the information seeking behaviour of online health information is limited. We conducted an experiment to investigate the information needs and behaviours of health information seekers. This paper reports on a model of behavioural patterns drawn on the experimental results, and implications for designing a better user experience for the exploration of online health information.

Categories and Subject Descriptors
H.1.2 [**Information Systems**]: User/Machine Systems – *Human information processing.*

General Terms
Design, Experimentation, Human Factors.

Keywords
Exploratory search, health, information seeking, human behaviour

1. INTRODUCTION
The Internet provides a variety of choices for consumer health information, from official health websites, private health service providers, personal blogs, to discussion forums. Studies have shown that lay-people look for health information online when they encounter health problems [1][2]. While we understand the demands of accessing health information on the Internet, there are few studies investigating the information needs and the characterising information seeking behaviours of these seekers. Our research aims to bridge this gap.

Search engines are the de-facto primary method of finding information on the web. This also applies to health seekers. However, the user experience in the health context is often not satisfactory due to a number of factors, including the nature of health information, the level of knowledge possessed by the seeker, and the skill of formulating search keywords [3][4][5][6].

When dealing with unfamiliar and unknown problems, or involving a task for which the goal remains unclear, the information seeking processes tends become more exploratory [7][8]. We argue that this exploratory search behaviour applies to health information seekers as well. In contrast to executing a search query and reading through the result, exploratory search involves a series of cognitive learning and query reformulation processes. A more complete picture of the knowledge domain is being built in this process. The exploratory process also implies the existence of both learning and investigative activities. Seekers end up knowing more information than they expected at the beginning. One of our goals is to design a system that better supports such exploration, which traditional search engines are not designed for.

In order to understand more deeply the effects on human behaviour of these health information needs, we ran interview sessions and observation experiments with 20 participants. During the observation, the participants displayed diverse behavioural patterns. Summarised from these patterns, we propose a model to describe the seeking behaviour in terms of research tactics and reading engagement. The findings indicate there exists a lack of features in both search engines and health websites to support health information seeking, which is essentially an exploratory search. We plan to build an experimental health website to address these problems in the next phase of our research.

2. LITERATURE REVIEW
In this section we draw on the literature to explain what we mean by the three phrases: online health information seeking, health information needs and exploratory search. These concepts are guiding the direction of this research.

2.1 Online Health Information Seeking
Health information seeking on the Internet is different from other types of searches in many ways. Lay-people usually have only limited knowledge in the medical domain [9], or face difficulties in utilising technical or medical language for searching [4][5]. On the other hand, dealing with health issues is stressful and uncertain and very likely to demonstrate a different information need [10].

A study showed 72% of U.S. Internet users have tried to access health information online [2]. Seekers look for a broad range of health information, which includes disease information, causes and treatments, diet information, health lifestyles, etc. [1][2][9][12][13][14][15], and in various stages of a health problem [16][17]. The diverse types of health information demanded reflect differences in information needs which result in different seeking behaviour.

MedIR 2014, July 11, 2014, Gold Coast, Australia.

2.2 Health Information Needs

Information needs arise when people realise their existing knowledge is inadequate to satisfy their goal. Finding information is an attempt to bridge a knowledge gap. A knowledge gap appears whenever people perceive there is not enough information in their minds and as a result they will start searching for information to fill the blanks [11]. This process is also known as a sense-making process [18].

Alzougool et al. investigated different information needs in the health context [10][19]. They propose that health information needs can be further classified into recognised and unrecognised needs. Cartright et al. [20] argued that health information seeking can be split into two partitions – evidence-based and hypothesis-directed. Through analysing queries in search engine sessions, they clustered the foci of searches in terms of causes, symptoms, remedies, or combinations of any of these. This could be useful to predict the information needs of exploratory health seekers.

2.3 Exploratory Search

Exploratory search involves learning and investigation in addition to lookup efforts, where the seeker interacts with information systems to retrieve a wider range of information [7]. Exploratory search can be found when the individual tries to address unfamiliar or unknown problems [21], as may be the case for health-related concerns. White and Roth add that people who are unfamiliar with the domain of their goals, or unsure about the ways to achieve their goals, or even unsure about their goals, will engage in exploratory search [8].

3. RESEARCH METHOD

Previous research about health information seeking behaviours does not capture well the actual patterns of the interactions among search engines and health websites. The aim of this experiment was to form an in-depth understanding of information needs and the patterns of related information seeking behaviours. We arranged sessions with individual participants in which we interviewed them about their experiences on health information seeking, and then gave them two tasks to carry out. We recorded the interviews and screen activities for further analysis.

The study was carried out from October to December of 2013. Posters were presented in various locations across the university for recruiting participants. E-mail invitations were sent out to encourage participation. Participants were also allowed to invite potential participants to our study through their connections. Recruitment continued until data saturation was achieved. This experiment was voluntary and no incentive was given to participants for the study.

Each study lasted about an hour and consisted of three sections: the first section was a semi-structured interview about their past experiences of finding health information on the web; participants were then given a computer to find online health information for two pre-defined tasks in the second part; the last section included another semi-structured interview to help researchers further understand how participants performed the tasks. We did not restrict the participants to use a specific website or instruct them to use a search engine. To avoid bias, we cleared the home page and all browsing history in the browser prior to each session. All interview content and screen activities were recorded.

The search tasks represented two different styles of common health scenarios. In one scenario, participants were asked to find information on how to care for a diabetic family member; in the other, participants were tasked with identifying information to append to a Wikipedia page on urination problems and their symptoms. The first scenario was designed with the aim of observing seeking behaviours for explicit recognised needs, whereas the second targeted for unrecognised needs with a more vague description of a health problem.

Interviews were transcribed and reduced to a number of codes iteratively [22]. Content that was relevant to health information seeking behaviours were organised into themes. Themes were derived with a thematic analysis approach [23]. Screen recordings were reviewed manually. A navigation graph was built for each participant to describe the web activity patterns in each session.

4. RESULTS

In total 20 participants completed our lab experiment (11 male; 9 female). In terms of identity, they comprised 8 students, 9 university staff and 3 external participants. The age distribution is listed in Table 1.

Table 1. Age Distribution of Participants

Age Group	N	Percentage
21-30	8	40%
31-40	6	30%
41-50	4	20%
Over 50	2	10%

4.1 Motivation for Exploratory Search

One of the themes emerging from the interviews regarded the motivation for exploratory search. Not every search about health topics is an instance of exploratory search, however, we have observed that both recognised and unrecognised information needs motivated people to perform exploratory search in our study.

Firstly, searches triggered by recognised needs are found to be more exploratory. For example, participants with health problems or those diagnosed with a certain illness. The need is also demanding if the issue is related to people's loved ones. They had a clearer mind about what information is helpful for the scenario, and possessed explicit information needs on aspects such as treatments and remedies. In this case, the recognised need mainly has the purpose of helping to understand the complete picture of the situation or getting more options for facing the health problem. The seeker is persistent in trying different ways to discover and read information as well.

Unrecognised needs were observed to stimulate exploratory search as well. People do not have a clear target, and therefore tend to approach different sources to make sense of the information – this is illustrative of exploratory behaviour during the search. Examples include people passively encountering contradictory messages, suspecting the validity of the information, or simply feeling curious about certain information. They seek additional sources without knowing what exactly is needed nor why it is needed. They usually become more open minded to the information obtained but are still cautious about it to avoid wrong information.

4.2 Behavioural Patterns

Participants were requested to perform two search tasks in our lab study sessions. For each task, we manually constructed a navigation graph to describe the pattern of the interaction among search engines, individual web pages and clicks on hyperlinks. 36

graphs were generated while 4 tasks not completed by participants due to time constraints on these lab sessions. From these graphs we identified four common patterns (Figure 1). During the search process of each task, the overall patterns of the individual's activities may contain one or more these common patterns shown in the figure.

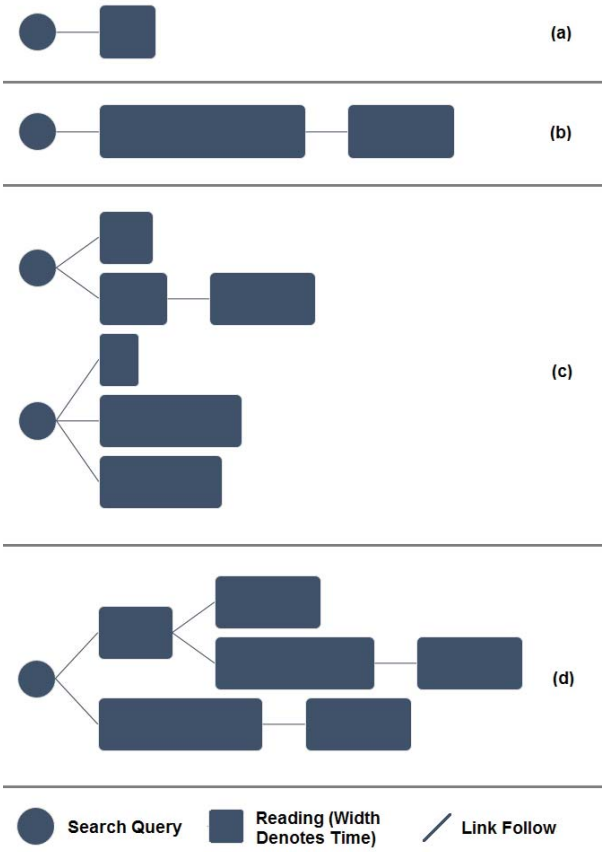

Figure 1. Typical Patterns of Health Exploratory Search

Figure 1(a) shows the simple pattern of executing a keyword query on a search engine, skimming one of the search results, and then finishing the information seeking process. This is common when the seeker just wishes to have some quick facts. The search stops once the information is found.

Figure 1(b) is similar to (a) in the small number of searches performed, but the seeker engages in longer and deeper readings. The seeker examines the web page returned by the search engine, also follows hyperlinks provided page-by-page and continues reading. The duration of reading is longer as the seeker digests and absorbs the information. The series of reading continues until no more valuable information is found.

Figure 1(c) presents not much reading but more query reformulations. The seeker in this case relies on the search engine to explore new information. Most of the time, he/she tries to skim through the search results and picks up only web pages that are relevant to the information need. Meanwhile the seeker will adjust the query keywords with the information read and submit a new query, if the overall search result is not satisfactory. The query reformulation particularly occurs when the user cannot discover new useful information with visible means (e.g. links) and feels the information in the website is exhausted.

Seekers in Figure 1(d) use hyperlinks to discover new information. They choose a small number of good websites (usually large and reputable) for examination. In these well-designed health websites, connections among pages are well-defined and hyperlinks are placed in a useful manner. Seekers can trace the related information through levels of hyperlinks easily and do not need to query the search engine as often as in other scenarios.

4.3 Model of Behavioural Outcome

From the perspective of enhancing user experience and suggesting design implications, conceptualising the information seeking behaviour is crucial. Drawn on the patterns in the previous section, we build up a model to abstract the behaviour outcome of online health information seeking behaviour (Figure 2).

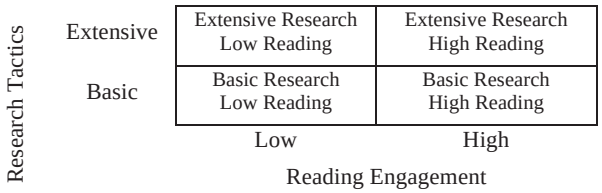

Figure 2. Model of Behavioural Outcome

This matrix presents a combination of seeker's research tactics and reading engagement. Research tactics represent the eagerness and motivation for finding out in-depth information. For more exploratory seekers, a wider range of information is needed, and thus the extensive tactics represent a greater effort to locate, filter, learn and discover other information within the current knowledge domain. Whereas less research effort will be put for the basic tactics, which often appear when looking for surfaced, easy-to-obtain and easy-to-read information.

Reading engagement measures the duration of reading and the intention of absorbing the information. Skimming and reading just the page summary fall into the group of low engagement while pursuing and digesting the information is considered as high engagement.

4.4 Guidelines for Design

Drawn on the abovementioned findings, we propose two areas of improvements for enhancing the user experience of exploratory health information seekers. These are (1) assisting the discovery of new information, and (2) adapting to users' reading needs.

The discovery of information within a website is important. In general, health websites collect many articles but we have noticed that seekers are not always able to reach all of them. This implies a problem of either users not knowing what they want, or that they cannot effectively use a search engine to explore. As seen in Figure 2, seekers with the basic research tactics stay within a single website rather than utilising a search engine for to look other sites. In this regard, a system that understands their information needs and recommends relevant information for further reading is preferable.

Figure 2 identifies a spectrum of reading engagement, suggesting that a health website needs to adjust to both low and high reading levels. In low engagement behaviour, users prefer to skim and quickly read through the articles to determine the usefulness before committing to a longer reading. In this case, an abstract or summary could be provided for their convenience. On the other hand, users with high reading engagement may prefer a design emphasising readability, such as font size, line spacing, section navigation, etc.

Extensive research seekers (the upper row in Figure 2) generally do not have major problems in finding the information. Corresponding to Figure 1 (c) and (d), they illustrate patterns of putting efforts to discover, locate and filter needed information. They also spend most of the time to judge the relevance of the materials, and seek alternatives if the information is not relevant. Both of the two design implications would be beneficial to this type of seekers.

5. CONCLUSION

This paper reports on a lab experiment with the aim to understand online health information seeking behaviour. From this study, we have identified that a variety of information needs that drive exploratory search, and a diversity in behavioural patterns. We have proposed a simple model using two factors, research tactics and reading engagement, that is useful for reflecting on user behaviour and starting to think about how we can design systems to provide better support for exploratory behaviours. These findings sight the directions we could work on to improve the user experience of health information seekers.

We will focus on designing a better environment for exploratory search in the health context. The next phase of this research is to build a testing health website with the goal on assisting the discovery of new information and enhancing reading engagement. A larger scale of user study will be launched to gather feedback and evaluate the new design elements in this new website.

6. ACKNOWLEDGEMENT

Patrick Pang is supported by the Australian Federal and Victoria State Governments and the Australian Research Council through the ICT Centre of Excellence program, National ICT Australia (NICTA).

7. REFERENCES

[1] Fox, S. and Jones, S. 2009. *The Social Life of Health Information*. Technical Report. Pew Internet & American Life Project.

[2] Fox, S. and Duggan, M. 2013. *Health Online 2013*. Technical Report. Pew Internet & American Life Project.

[3] Zhang, Y. 2011. Exploring a web space for consumer health information: implications for design. In *Proceedings of iConference 2011* (Seattle, WA).

[4] Keselman, A., Browne, A. C. and Kaufman, D. R. 2008. Consumer health information seeking as hypothesis testing. *JAMIA*. 15, 4 (Jul. 2008), 484-495.

[5] Chapman, K., Abraham, C., Jenkins, V. and Fallowfield, L. 2003. Lay understanding of terms used in cancer consultations. *Psycho-oncology*. 12, 6 (Sep. 2003), 557-566.

[6] Luo, G., Tang, C., Yang, H. and Wei, X. 2008. MedSearch: A specialized search engine for medical information retrieval. In *Proceedings of 17th ACM Conference on Information and Knowledge Management* (Napa Valley, CA, October 26-30, 2008). CIKM '08.

[7] Marchionini, G. 2006. Exploratory search: from finding to understanding. *Communications of the ACM*. 49, 4 (Jun. 2006), 41-46.

[8] White, R. and Roth, R. 2009. *Exploratory Search: Beyond the Query-Response Paradigm*. Morgan and Claypool, San Rafael, CA.

[9] Zhang, Y., Fu, W.-T. 2011. Designing consumer health information systems: what do user-generated questions tell us? In *Proceedings of the FAC 2011, HCII 2011, LNAI 6780*.

[10] Alzougool, B., Chang, S. and Gray, K. 2013. The nature and constitution of informal carers' information needs: what you don't know you need is as important as what you want to know. *information Research*. 18, 1 (Mar. 2013).

[11] Case, D. O. 2002. *Looking for Information: A Survey of Research on Information Seeking, Needs, and Behaviour*. Academic Press, Lexington.

[12] Andreassen, H. K., Bujnowska-Fedak, M. M., Chronaki, C. E., Dumitru, R. C., Pudule, I., Santana, S., Voss, H. and Wynn, R. 2007. European citizens' use of E-health services: a study of seven countries. *BMC public health*. 7, 53 (Apr. 2007), 1-7.

[13] Bessell, T. L., Silagy, C. A., Anderson, J. N., Hiller, J. E. and Sansom, L. N. 2002. Prevalence of South Australia's online health seekers. *Australian and New Zealand J. of Public Health*. 26 (Mar. 2002), 170-173.

[14] Johnson, J. D. and Meischke, H. 1991. Women's preferences for cancer information from specific communication channels. *The American Behavioral Scientist*. 34, 6 (Jul. 1991), 742.

[15] Nicholas, D., Huntington, P., Gunter, B., Withey, R. and Russell, C. 2003. The British and their use of the web for health information and advice: a survey. *Aslib Proceedings*. 55, 5, 261-276.

[16] Rutten, L. J. F., Arora, N. K., Bakos, A. D., Aziz, N. and Rowland, J. 2005. Information needs and sources of information among cancer patients: a systematic review of research (1980–2003). *Patient Education and Counseling*, 57(2005), 250–261.

[17] Ofran, Y., Paltiel, O., Pelleg, D., Rowe, J. M. and Yom-Tov, E. 2012. Patterns of information-seeking for cancer on the Internet: An analysis of real world data. *PLoS ONE*, 7(9), e45921.

[18] Wilson, T. D. 1999. Models in information behaviour research. *Journal of Documentation*, 55, 3, 249-270.

[19] Alzougool, B., Chang, S. and Gray, K. 2008. Towards a comprehensive understanding of health information needs. *electronic Journal of Health Informatics*, 3, 2.

[20] Cartright M.-A., White, R. W., and Horvitz, E. 2011. Intentions and attention in exploratory health search. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR'11. 65-74.

[21] Pearce, J., Chang, S., Kennedy, G., Ely, R. B. W. and Ainley, M. 2012. Search and explore: more than one way to find what you want. In *Proceedings of the 2012 Australian Computer-Human Interaction Conference* (Melbourne, Victoria, Australia, November 26-30, 2012). OzCHI '12.

[22] Creswell, J. W. 2002. *Educational Research: Planning, Conducting, and Evaluating Quantitative and Qualitative Research*. Merrill Prentice Hall, Upper Saddle River, NJ.

[23] Braun, V. and Clarke, V. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3, 2 (Jan. 2006), 77-101.

Evaluation of Coreference Resolution for Biomedical Text

Miji Choi
[1]The University of Melbourne
Melbourne, Australia
[2]National ICT Australia
jooc1@student.unimelb.edu.au

Karin Verspoor
The University of Melbourne
Melbourne, Australia
karin.verspoor@unimelb.edu.au

Justin Zobel
The University of Melbourne
Melbourne, Australia
jzobel@unimelb.edu.au

ABSTRACT
The accuracy of document processing activities such as retrieval or event extraction can be improved by resolution of lexical ambiguities. In this brief paper we investigate coreference resolution in biomedical texts, reporting on an experiment that shows the benefit of domain-specific knowledge. Comparison of a state-of-the-art general system with a purpose-built system shows that the latter is a dramatic improvement.

Categories and Subject Descriptors
Computing methodologies:: artificial intelligence:: natural language processing:: information extraction, phonology/ morphology; Applied computing:: life and medical science:: health informatics.

General Terms
Algorithms, Performance, Reliability.

Keywords
Coreference resolution, domain-specific knowledge, named entity recognition.

1. INTRODUCTION
The peer-reviewed scientific literature is a vast repository of authoritative knowledge. The life sciences literature is the basis of biomedical research and clinical practice, and must be searchable to be of value. However, with around 40,000 new journal papers every month, manual discovery or annotation is infeasible, and thus it is critical that document processing techniques be robust and accurate, to enable not only conventional search, but automated discovery and assessment of knowledge such as interacting relationships (events and facts) between biomolecules such as proteins, genes, chemical compounds and drugs. Biological molecular pathways, for example, integrated with knowledge of relevant protein-protein interactions, or chemical reactions, are used to understand complex biological processes that could explain specific health conditions in human body in biomedical and pharmaceutical research.

A particular challenge is the need for lexical ambiguity resolution [1]. Lexical ambiguity is a general problem for text processing – such as for search or for event extraction – but is particularly acute in this domain, which has a vast but inconsistent technical lexicon; the domain also presents particular opportunities, because many technical terms are constructed in accordance with a set of highly standardized rules. Thus while there are particular kinds of ambiguity (genes and proteins may share names, for example) there are also deductions that can be made from name structure (for example, that a certain name must be a chemical).

A key obstacle is the low detection reliability of hidden or complex mentions of entities involving coreference expressions in natural language texts [2, 3]. Thus, coreference resolution is an essential task in information extraction, because it can automatically provide links between entities, and as well can facilitate better indexing for medical information search with rich semantic information.

For example, the following passage includes an interacting relation; the *binding* event between the anaphoric mention *the protein* and a cell entity *CD40* is implied in the text. The mention *the protein* refers to the specific protein name, *TRAF2*, previously mentioned in the same discourse.

> *… The phosphorylation appears to be related to the signalling events that are activated by TRAF2 under these circumstances, since two non-functional mutants were found to be phosphorylated significantly less than the wild-type protein. Furthermore, the phosphorylation status of TRAF2 had significant effects on the ability of* ***the protein to bind to CD40****, as evidenced by our observations …*

Such anaphoric mentions, or pronouns in texts, are mostly ignored by event extraction systems, and are not considered as term occurrences in information retrieval systems. In this brief paper, we report an initial investigation of the challenges of biomedical coreference resolution, test an existing general domain coreference resolution system on biomedical texts, and demonstrate that domain-specific knowledge can be helpful for coreference resolution for the biomedical domain.

2. EXPERIMENT
To evaluate the important of domain-specific knowledge, we compare an existing coreference resolution system, TEES, that uses a domain-specific named entity recognition (NER) module with an existing general system, CoreNLP, that does not use a domain-specific NER. The aim is to explore how domain-specific information impacts on performance for coreference resolution involving protein and gene entities. The TEES system, which includes a biomedical domain-specific NER component for protein and gene mentions [4], and the Stanford CoreNLP system, which uses syntactic and discourse information but no NER outputs [5], are evaluated on a domain-specific annotated corpus.

MedIR2014, July 11, 2014, Gold Coast, Australia.

2.1 Data Sets

We use the training dataset from the Protein Coreference Shared task at BioNLP 2011 [2] for our evaluation of existing coreference resolution systems. The annotated corpus includes 2,313 coreference relations, which are pairs of anaphors and antecedents related to protein and gene entities, from 800 Pubmed journal abstracts. As shown in Table 1, this gold standard dataset consists of coreference relations involving relative pronouns such as *which*, *that*, or *who*, or pronouns such as *it*, *its*, or *they*. Among 2,313 coreference relations, 560 relations embed one or more specific protein and gene name.

Table 1. Statistics of the annotated corpus at the coreference relation level

Anaphor	Relative pronoun	1,174 (51%)
	Pronoun	754 (32%)
	Definite Noun Phrase	346 (15%)
	Indefinite Noun Phrase	11 (0.5%)
	Proper Noun	22 (1%)
	Unclassified	6
Antecedent	Including protein/gene	560
	Including conjunction	217
	Cross-sentence	389
	Identical relation	43
	Head-word match	254

2.2 Results

Performance for identification of coreference mentions and relations of each system evaluated on the annotated corpus is compared in Table 2. The Stanford system achieved low performance with F-score 12% and 2% for the detection of coreference mentions and relations respectively, and produced a greater number of detected mentions, while the TEES system achieved better performance with F-score 69% and 37% for coreference mention and relation levels respectively, but produced smaller number of detections, which reduced system recall. Both systems demonstrate huge reduction in detection of coreference relations from the mention detection with the number of exact matched 1,006 at the mention level to 112 by the Stanford system, as well as from 2,466 to 546 by the TEES system.

Table 2. Results of evaluation of existing systems on the annotated corpus

	Stanford		TEES	
	Mention	Relation	Mention	Relation
Gold corpus	4,367	2,313	4,367	2,313
System detected	12,848	7,387	2,796	707
Exact match	1,006	112	2,466	564
Precision	0.08	0.02	0.88	0.80
Recall	0.23	0.05	0.56	0.24
F-score	0.12	0.02	0.69	0.37

Our investigation of low performance by each system at the coreference relation level is analysed in detail in Figure 1.

	Stanford			TEES			
	Cross-sentence	Internal-sentence	Including protein	Cross-sentence	Internal-sentence	Including protein	
Relative pronoun	TP 0 FP 0	TP 1 FP 2	TP 0 FP 1	TP 0 FP 0	TP 393 FP 86	TP 116 FP 27	D
Pronoun	TP 7 FP 675	TP 62 FP 302	TP 28 FP 197	TP 0 FP 0	TP 162 FP 47	TP 37 FP 15	B
Definite noun phrase	TP 35 FP 1183	TP 7 FP 194	TP 10 FP 483	TP 0 FP 0	TP 7 FP 3	TP 2 FP 1	E
Indefinite noun phrase	TP 0 FP 62	TP 0 FP 81	TP 0 FP 49	TP 0 FP 0	TP 1 FP 2	TP 0 FP 0	
Unclassified	TP 0 FP 4129	TP 0 FP 650	TP 0 FP 1187	TP 0 FP 0	TP 1 FP 5	TP 0 FP 3	
			A	C		A	

Figure 1. Analysis of performance of existing systems comparing to the annotated corpus

Several factors such as lack of domain-specific knowledge (A), bias towards selection of closest candidate of antecedent (B), limiting analysis to within-sentence relations (C), syntactic parsing error (D), and disregard of definite noun phrase (E) have been observed. The main cause, lack of domain-specific knowledge, is explored below.

The annotated corpus contains 560 coreference relations, where anaphoric mentions refer to protein or gene entities previously mentioned in a text. For those coreference relations, the TEES system outperformed the Stanford system by identifying 155 true positives – far more than the 38 identified by the Stanford system, as shown in Table 3.

Table 3. Result of performance of existing systems for coreference relations involving protein names

	Output		Precision	Recall	F-score
Stanford	TP	38	0.02	0.07	0.03
	FP	1732			
TEES	TP	155	0.77	0.28	0.41
	FP	46			

The Stanford system also produces a large number of false positives. Even though half of the false positives are relations where anaphors are unclassified, the system links coreference relations where an anaphor and an antecedent are identical, or have a common head word (the main noun of the phrase). This is because coreference resolution systems in general domains aim to identify all mentions that refer to the same entity in a text, rather than to resolve only specifically anaphoric mentions. Considering those anaphoric mentions, inspection of individual instances (as illustrated in Figure 2) strongly suggests that lack of domain-specific knowledge is the main cause of failure.

On the other hand, the TEES system achieved 77% precision, but still only 28% recall. The main reason for the low recall is that the system is limited to identification of coreference relations where anaphors and antecedents corefer within a single sentence. Even though anaphoric coreference mentions mostly link to their antecedents across sentences, the system still identified 155 correct coreference relations by taking advantage of domain-specific information provided through recognition of proteins.

Figure 2 demonstrates how the process of NER in the biomedical domain helps to determine correct coreference relations. In the

example, the anaphoric mention *the protein* is correctly identified as referring to *TRAF2* by the TEES system, but the Stanford System links it to the incorrect antecedent *the wild-type protein*.

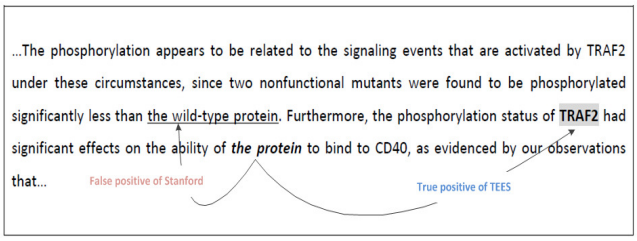

Figure 2. Example of a coreference relation involving a protein entity, and results of coreference resolution performed by both the TEES and the Stanford systems

3. CONCLUSIONS

In this study, we have explored how domain-specific knowledge can be helpful for resolving coreferring expressions in the biomedical domain. The performance difference between a system using a domain-specific NER approach and a general system is substantial. In detailed analysis of individual cases of failure (not reported here) we have observed that the domain knowledge, rather than variation in methods, is the main explanation for the success of the domain-specific approach.

4. ACKNOWLEDGMENTS

This work was supported by the University of Melbourne, and by the Australian Federal and Victorian State governments and the Australian Research Council through the ICT Centre of Excellence program, National ICT Australia (NICTA).

5. REFERENCES

[1] Krovetz, R. *Homonymy and polysemy in information retrieval*. Association for Computational Linguistics, 1997.

[2] Nguyen, N., Kim, J.-D. and Tsujii, J. i. *Overview of the protein coreference task in BioNLP shared task 2011*. Association for Computational Linguistics, 2011.

[3] Miwa, M., Sætre, R., Kim, J.-D. and Tsujii, J. i. Event extraction with complex event classification using rich features. *Journal of bioinformatics and computational biology*, 8, 01 2010), 131-146.

[4] Björne, J. and Salakoski, T. *Generalizing biomedical event extraction*. Association for Computational Linguistics, 2011.

[5] Lee, H., Peirsman, Y., Chang, A., Chambers, N., Surdeanu, M. and Jurafsky, D. Stanford's multi-pass sieve coreference resolution system at the CoNLL-2011 shared task. Association for Computational Linguistics, 2011.

Retrieving Attitudes: Sentiment Analysis from Clinical Narratives

Yihan Deng
ICCAS
University of Leipzig
Semmelweissstr. 14
Leipzig, Germany

Matthaeus Stoehr
ENT Clinic
University Hospital Leipzig
Liebigstr. 10
Leipzig, Germany

Kerstin Denecke
ICCAS
University of Leipzig
Semmelweissstr. 14
Leipzig, Germany

{name.surname}@iccas.de

ABSTRACT

Physicians and nurses express their judgments and observations towards a patient's health status in clinical narratives. Thus, their judgments are explicitly or implicitly included in patient records. To get impressions on the current health situation of a patient or on changes in the status, analysis and retrieval of this subjective content is crucial. In this paper, we approach this question as sentiment analysis problem and analyze the feasibility of assessing these judgments in clinical text by means of general sentiment analysis methods. Specifically, the word usage in clinical narratives and in a general text corpus is compared. The linguistic characteristics of judgments in clinical narratives are collected. Besides, the requirements for sentiment analysis and retrieval from clinical narratives are derived.

Categories and Subject Descriptors

H.3 [**INFORMATION STORAGE AND RETRIEVAL**]: Content Analysis and Indexing

Keywords

Clinical text mining, Sentiment analysis

1. INTRODUCTION

Sentiment analysis deals with determining the sentiment with respect to a specific topic expressed in natural language text. So far, the development of sentiment analysis methods concentrated on processing very opinionated, subjective texts such as customer reviews [3, 4]. Clearly, sentiment in clinical documents differs from sentiment in user-generated content or other text types. With the term *sentiment* we refer to information on the health status, or on the outcome of a medical treatment or change / seriousness of a symptom (e.g. *serious pain*) or the certainty of an observation. The work presented in this paper intends to get a more complete view on the facets of *sentiment* in clinical texts. With the development of the principles of evidence-based medicine [6] and digital patient modeling [1], the observations and judgments expressed in clinical narratives will play a crucial role for the clinical decision process.

Consider the following scenario: During the daily ward round, a physician is making observations with respect to the health status of a patient (e.g. symptoms improved). The patient describes his personal experiences on the symptoms such as the degree of pain. All this information reflects the individual health status and is documented in clinical notes. Retrieving, analyzing and aggregating this information over time can support the treatment decisions and allows a physician to quickly get an overview on the health status. Another application example is retrieving attitudes from clinical documents which can support assessing the outcome of treatments. In this way, labor-intensive user studies for treatment or medication evaluation can be facilitated.

For processing clinical narratives in the last years, effective algorithms in particular for named entity recognition and relation extraction [2] have been developed. Based on recognized entities and relations between entities, sentiments expressed in medical narratives can now be analyzed to offer an upper-level text understanding. Further, a corresponding retrieval of judgments or sentiments can be realized. However, sentiments, opinions and intentions expressed in clinical narratives have not been well exploited yet. In this paper, we start analyzing the sentiment expressions used in clinical texts through a linguistic comparison with a non-medical, subjective text corpus.

Conventional methods for sentiment analysis have been developed for processing subjective on-line documents such as weblogs and forums. In this paper, our goal is to analyze the applicability of such methods for sentiment analysis in clinical narratives. We will identify necessary extensions of existing methods and come up with the requirement of sentiment in clinical narratives. To this end, we will first compare two types of medical narratives (radiology report and nurse letter) with a weblog data set. The lexical and linguistic differences will be presented. Afterwards, we will apply a general subjectivity lexicon to medical narratives using dictionary-based methods. Sources of errors of this simple sentiment recognition approach will be discussed. The following research questions will be addressed:

1. In comparison with user generation content, which lexical characteristics do clinical narratives have?

MedIR July 11, 2014, Gold Coast, Australia

2. What characterizes sentiments in clinical narratives?

3. Can existing methods for sentiment analysis be applied? Which adaptations are necessary?

2. SENTIMENT ANALYSIS IN THE MEDICAL DOMAIN

To our best knowledge, few work considered sentiment analysis in medical texts: Xia et al. [9] have indicated that sentiments are topics-related. Their approach to sentiment analysis starts with a standard topic classifier based on topic labels. In the second step, special classifiers are initialized to detect the polarity for each topic. The multi-step classification method has earned a nearly 10% improvement of F1 measure in comparison with the single-step approach. Niu et al. consider sentiment analysis in biomedical literature [5]. They exploit a supervised method to classify the polarity at sentence level. The linguistic features such as uni-grams, bi-grams and negations are employed. The medical terms are merely replaced by their semantic category. The category information and context information are derived from the Unified Medical Language System (UMLS[1]). The combination of linguistic features and domain-specific knowledge have improved the accuracy of the algorithm.
In summary, existing methods for sentiment analysis in the medical domain focus on processing biomedical literature and patient-generated text. The clinical text which is used to record the activities and judgments of health care workers has not yet been analyzed. Moreover, the existing approaches and definitions of sentiment in the medical domain are derived from general sentiment analysis for Web 2.0 media. Clinical context and medical knowledge have not been used thoroughly besides some category meta data derived from the UMLS [7, 5]. We expect that due to different expressions and the more objective way of writing in the clinical narratives, the conventional sentiment analysis methods need to be adapted to cope with the clinical context. We will concentrate on that particular text material.

3. METHODOLOGY

3.1 Text Material

In order to analyze the differences between the language in clinical narratives and general texts from the Internet, 200 nurse letters and 200 radiology reports from "MIMIC II Database[2]" have been chosen as corpus. These documents form the domain-specific data source in our assessment. For comparison reasons, we additionally consider 200 technical interviews downloaded from the website Slashdot[3]. We have chosen that particular dataset since it belongs to the category of user-generated, subjective content. Given the technical topics, we however expect a certain similarity, mainly an objectivity as it occurs in clinical narratives.
Nurse Letter: A nurse letter is part of a patient record, and is written by nurses on duty. Its content reflects the situation of the patient and the feedback to the ongoing treatment. It is written in a relatively subjective manner. Acronyms and typos appear very often in nurse letters.

[1]http://www.nlm.nih.gov/research/umls/, accessed: 20.04.2014
[2]http://www.physionet.org/, accessed 20.04.2014
[3]http://slashdot.org, accessed 20.04.2014

Radiology Report: A radiological report is mainly used to inform the treating physicians about the findings in an radiological examination. It starts usually with a medical history, which is followed by a description of the region of interest and questions for the examinations. The texts contain many judgments and observations as observed in the examination.
Slashdot Interviews: Slashdot is a technology-related weblog, which covers different technical topics. The users express their opinions on certain topics. We chose the technical interviews as benchmark instead of movie or product review, since technical interviews contain also a relatively large amount of terminologies.

3.2 Linguistic and Sentiment Analysis of the Data Sets

Apparently, the three text sources are different in terms of terminology usage and content. The interview corpus is typical user generated content. We expect that the corpus will contain a relatively large amount of sentiment terms and subjective expressions, while the clinical narratives are written in a more objective way. Less opinionated terms and rather more clinical terminology are expected. However, the question is whether the terminology and word usage is really distributed as expected. To what extent do the corpora differ with respect to linguistic characteristics? Recalling our initial research questions, we need to answer whether existing sentiment lexicons can provide the basis for analyzing judgments and sentiments in clinical narratives. In order to address these questions, an extraction pipeline has been built to obtain part of speeches and sentiment terms from the texts and to determine their occurrence frequency. The Penn Tree POS-tagger[4] and the SL sentiment lexicon [8] (contains 8,221 single-term subjective expressions) have been exploited for this purpose. The punctuation, numbers and stop words were also extracted and their proportions were calculated.

After analyzing the linguistic composition of the data sets, we want to study the applicability of a dictionary-based sentiment analysis approach on clinical narratives. Potential limitations of the approach when applied to medical narratives will be identified. For this purpose, we have created an experiment pipeline in KNIME[5]. Two dictionary taggers were applied to recognize positive and negative terms in the text respectively. A voting algorithm is applied to calculate the polarity for each document. It is based on the number of positive and negative occurrences and handles negations. Although it is only a simple approach, it is a direct method to evaluate the compatibility between the subjectivity lexicon and clinical narratives. The SL sentiment lexicon from Wilson et al. [8] is used by the dictionary tagger. It comprises a large amount of adjectives, adverbs, but also nouns and verbs expressing sentiments. For evaluation purposes, the three corpora were annotated with an overall document polarity at document level by one physician from our university hospital.

4. RESULTS AND DISCUSSION

4.1 Results of the Linguistic Analysis

[4]http://www.cis.upenn.edu/ treebank/, accessed 20.04.2014
[5]http://www.knime.org/, accessed 20.4.2014

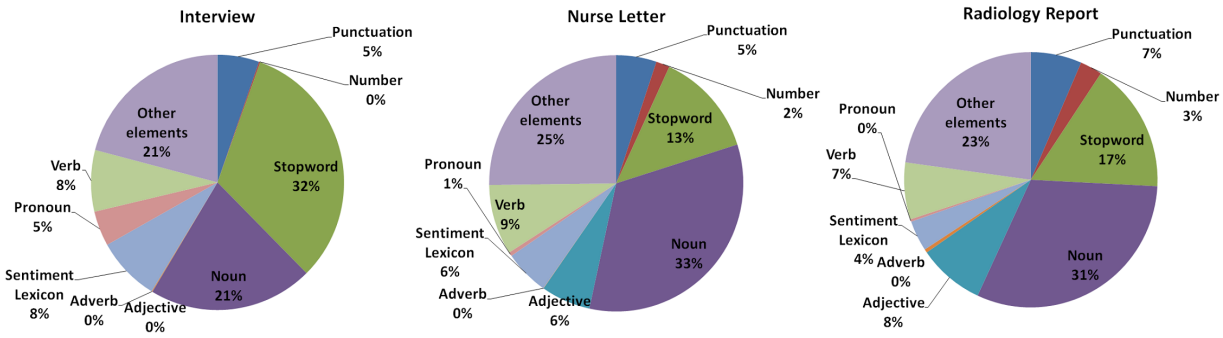

Figure 1: Result of the Linguistic Analysis

In Figure 1, the proportions of punctuations, numbers, stop words, nouns, pronouns, adjectives, adverbs as well as the sentiment terms are illustrated. Part of speeches of terms that matched with the sentiment lexicon have not been considered. The result has partially confirmed our expectation.

Sentiment Terms: According to the results, the normal interview corpus contains the highest proportion of sentiment terms with 8%, while the radiology reports contain 4% and nurse letters 6% sentiment terms. These results have approved our observation that nurse letters are written more subjectively in comparison to radiology reports, but they are still more objective than the interviews. The sentiments expressed in nurse letters are normally implicit and appear with the description of patient's health status, or the social records for the visitors of the patients. Opinionated terms and expressions such as suspicion, negation, approval or recommendations can be found in radiology reports mainly in the conclusion section or impression part at the end of the whole report.

Number: Numbers are one of the most important elements in clinical reports, where they are mainly used to represent the dose of medications, the size of a tumor or the frequency of a treatment, etc. In our clinical data sets, numbers comprise between 2% and 3% of the words or characters. In contrast, in the interviews almost no numbers occur, since the discussions in weblogs are more likely to use simple, colloquial vocabulary to present the personal attitudes and preferences.

Stop Word: The nurse letters and radiology reports contain 13% and 17% stop words respectively. In contrast, the percentage of stop words in the interview corpus is with 32% significantly higher, which shows that the clinical documents are clearly written in a concise way, focusing on facts.

Nouns and Pronouns: What noteworthy is, the percentage of nouns in radiology reports (31%) and nurse letter (33%) is clearly higher than the percentage of nouns in interviews (21%), while the percentage of pronouns in the interviews (4%) is notably higher than in the radiology report (0%) and nurse letters (1%). The reason is that in medical facts are described in clinical narratives using nouns from medical terminologies (e.g. names of diseases, symptoms , medications). In contrast, the interviews contain more subjective terms and use a large amount of first person expressions to express the ideas and opinions of individuals.

Adjective and Adverb: Another interesting finding is that the clinical narratives contain a substantial amount of adjectives (6-8% of the terms) that are not included in the SL sentiment lexicon. In contrast, all adjectives in the interview corpus matched with the sentiment lexicon. The additional adjectives in clinical narratives are mainly related to body locations, such as "left" side, "right" side, "vertical", "dorsal", "cervical". They express neither emotion nor attitude but anatomical concepts and relative locations in the body. In summary, the nurse letters show a relatively higher linguistic similarity to technical interviews than radiology reports. They are to a certain extent more subjectively written than radiology reports. The large amount of the medical terms (noun, adjective) describe the status of a patient. They reflect the attitudes of physicians. Thus, the implicit clinical events may influence the polarity outcome of a clinical report as well. Consequently, the implicit clinical events and evidences are expected to be relevant to understand and interpret the status of the patient.

Types	Accu(overall)	F1(Bad)	F1(Neutral)	F1(Good)
Interviews	0.696	0.754	0.367	0.735
Nurse Letter	0.420	0.437	0.216	0.503
Radiology Report	0.446	0.297	0.080	0.559

Table 1: Sentiment Analysis Results: Accuracy and F1 measure for three text types

4.2 Results of the Sentiment Analysis

The automatically retrieved polarity for the texts were compared to the manual annotation done by clinical experts. The overall accuracy and F1 measure for the three text types is shown in Table 4.1: Accuracy is the proportion of true results in the population. The sentiment analysis of interviews leads to an acceptable accuracy of 69.6%. The results for nurse letters and radiology reports have merely achieved the accuracies of 42% and 44% respectively. This shows that existing methods need to be adapted when processing these texts and that **sentiment** is different. Furthermore, the F1 measure for positive texts (F1 good) is significantly higher for the clinical texts than for negative (F1 bad) texts. A manual assessment showed that the positive sentiments or outcomes are described in an explicit way, e.g., by phrases such as the "patient slept well, the treatment has a satisfactory result" or "the tube has been placed successfully". For negative clinical events, the nurse and physician were more

likely to express the status of patient in a careful and cautious manner, e.g. by phrase such as "some situation cannot be excluded or need further pathological investigation". The radiology reports are more likely to exclude or confirm the occurrences of certain clinical events rather than to give a final diagnosis. In addition, the recognition of neutral situations is difficult, since the judgment of neutral outcome depends on the recognition of positive and negative terms. However, neutral clinical outcomes in real world are probably not objectively expressed. Some surgical result may only show moderate effect, but it may turn out to be an insignificant outcome in nurse letters or might even produce some negative feedbacks. During the annotation, our physician tended to give more positive and negative judgments to the reports rather than neutral ones, since the determination of "neutral" needs more context and reference, which is not that easy to obtain without knowing the complete patient history.

5. CONCLUSION AND FUTURE WORK

In this paper, we have studied the linguistic characteristics of clinical narratives compared to a web data set and analyzed the feasibility of a simple sentiment analysis approach on clinical narratives. The results provide important insights to understand **sentiment** in clinical narratives and to continue with developing corresponding analysis methods. The initial three research questions raised in Section 1 can be answered.

1. The linguistic analysis showed that clinical narratives contain a moderate amount of sentiment terms. In contrast to the web data set, more numbers, medical terms (nouns), location-related adjectives are exploited and less stop words, and less pronouns are included. This composition and word usage reflects the objectivity and preciseness of the clinical writing style.

2. By analyzing the clinical documents, we learned more about the nature of sentiment in clinical narratives. Sentiment can concern the general health status of a patient, the outcome of a treatment or of a specific medical condition or can concern uncertainty of an observation. **Good, bad** or **positive** and **negative** is manifested in status changes, e.g. an improvement or worsening of a certain medical or physical condition or the success or failure of a treatment. Sentiment can be seen as health status of a patient: The patient's health status can be **good, bad** or **normal** at some point in time, expressed either implicitly or explicitly. By analyzing that health status over time, improvements or worsening in the status can be recognized. An implicit description of a health status concerns the mentioning of critical symptoms (e.g. serious pain, extreme weight loss, high blood pressure). A explicit description of the health status is reflected through phrases such as "the patient recovered well" or "normal". Sentiment in clinical texts can be the outcome of a treatment or the impact of a specific medical condition, i.e. whether the condition improved or worsens which allows to draw conclusions on the effect or outcome of a treatment (positive/negative outcome). The phrase "blood sugar decreased" could express a positive or negative change depending on the previous state. A decrease of blood pressure can be good when it was too high before. This also shows that for interpreting the detected sentiment, the context need to be considered. Further, sentiment can be seen as presence, change in or certainty of a medical condition. I.e. a medical condition can exist, improve, worsen, be certain or uncertain. The treatment outcome can be positive, negative (e.g. surgery was successful or failed), neutral or a treatment can have no outcome.

3. A simple method for sentiment analysis is not well suited to analyze sentiment in clinical narratives. Sentiment in clinical texts differs significantly from sentiment in general texts. In particular, implicit sentiments need to be detected. An adapted annotation scheme should be defined with the help of physicians. New features for sentiment analysis need to be collected for gathering these subjective sentiments.

In the short term, we will develop a sentiment lexicon specific for the medical domain. It will define a scheme for analyzing and retrieving implicit sentiments and attitudes expressed in clinical texts. The kind of influence and degree of influence of a symptom to the health status will be defined. This lexicon or ontology will be exploited for developing a more comprehensive sentiment analysis algorithm.

6. REFERENCES

[1] K. Denecke. Model-based Decision Support: Requirements and Future for its Application in Surgery. *Biomedical Engineering*, 58(1), 2013.

[2] C. Friedman, T. C. Rindflesch, and M. Corn. Natural language processing: State of the art and prospects for significant progress, a workshop sponsored by the national library of medicine. *Journal of Biomedical Informatics*, 46(5):765–773, 2013.

[3] M. Hu and B. Liu. Mining and summarizing customer reviews. In *Proc. Tenth ACM SIGKDD*, KDD '04, pages 168–177, New York, NY, USA, 2004. ACM.

[4] B. Liu. *Sentiment Analysis and Opinion Mining.* Synthesis Lectures on HLT. Morgan & Claypool Publishers, 2012.

[5] Y. Niu, X. Zhu, J. Li, and G. Hirst. Analysis of polarity information in medical text. In *In: Proc of the AMIA 2005 Annual Symposium*, pages 570–574, 2005.

[6] D. L. Sackett, W. M. Rosenberg, J. Gray, R. B. Haynes, and W. S. Richardson. Evidence based medicine: What it is and what it isn't. *BMJ*, 312(7023):71–72, 1996.

[7] A. Sarker, D. Molla, and C. Paris. Outcome polarity identification of medical papers. In *Proceedings of the Australasian Language Technology Association Workshop 2011*, pages 105–114, Canberra, Australia, December 2011.

[8] T. Wilson, J. Wiebe, and P. Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proc of HLT '05*, HLT '05, pages 347–354, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.

[9] L. Xia, A. L. Gentile, J. Munro, and J. Iria. Improving patient opinion mining through multi-step classification. In *TSD*, pages 70–76, 2009.

Why Assessing Relevance in Medical IR is Demanding

Bevan Koopman
Australian e-Health Research Centre, CSIRO
Brisbane, Australia
bevan.koopman@csiro.au

Guido Zuccon
Queensland University of Technology
Brisbane, Australia
g.zuccon@qut.edu.au

ABSTRACT

This study investigates if and why assessing relevance of clinical records for a clinical retrieval task is cognitively demanding. Previous research has highlighted the challenges and issues information retrieval systems are faced with when determining the relevance of documents in this domain, e.g., the vocabulary mismatch problem. Determining if this assessment imposes cognitive load on human assessors, and why this is the case, may shed lights on what are the (cognitive) processes that assessors use for determining document relevance (in this domain). High cognitive load may impair the ability of the user to make accurate relevance judgements and hence the design of IR mechanisms may need to take this into account in order to reduce the load.

Categories and Subject Descriptors: H.3 [Information Storage and Retrieval]: H.3.3 Information Search and Retrieval
General Terms: Experimentation.

1. INTRODUCTION

The collection of relevance assessments is important for information retrieval (IR) systems evaluation. Relevance is a complex notion: subjective to the person performing the assessment, dependent on contextual factors and often acting on multiple dimensions (i.e., factors like opinion, readability and trustworthiness may influence a relevance judgement) [3]. To the best of our knowledge, however, there has been little or no work that investigates if and why it is cognitively demanding for assessors to judge relevance.

In this paper, we aim to determine: (i) if assessing document relevance is demanding; if so (ii) what are the indicators of a demanding assessment; and (iii) what are the reasons behind an assessment being demanding or not. Toward these aims, we focus on medical IR, and more specifically on the task of finding patients suitable to clinical trials, i.e., the task modelled in the TREC Medical Records Track (MedTrack) [5]. It has been shown that this is, in general, a difficult task for IR systems due to factors like vocabulary and granularity mismatch, conceptual implication, and inferences of similarity [1]. However, no previous work has explored whether this also applies for humans, and whether assessing the relevance of health records for this task is cognitively demanding (indeed, difficult) for expert assessors.

MedIR July 11, 2014, Gold Coast, Australia.
ACM SIGIR.

Given the familiarity that medical experts have with medical documents, one may posit that the task of assessing relevance in these documents is not demanding for experts. On the contrary, our quantitative and qualitative analysis of a relevance assessment exercise, performed by four experts, revealed that assessing relevance in the medical domain is often demanding: assessments required substantial time to be formed, implying a substantial cognitive load on the assessors. Given this result, we explore and validate a number of factors associated to both queries and documents that contribute to the difficulty of the assessment task, revealing why this task is demanding.

2. EXPERIMENTAL DESIGN

We used data gathered from a previous relevance assessment task [1]. In this previous study, four medical professionals were asked to judge clinical documents taken from the TREC MedTrack collection [5]. As we used data from an existing study not explicitly designed to fully answer the research questions of this paper, we are constrained by the data captured in the previous study. Nevertheless, a number of insights into how demanding assessment are can be derived.

The original TREC MedTrack queries were used and a total of 1030 documents were assessed.[1] To collect assessments, the Relevation! judging system was used [2]. Queries were divided between the four assessors with each query being fully judged by only one assessor. Each assessor also completed two control queries to familiarise themselves with the task. As all assessors completed the same control queries, these were used to determine inter-coder agreement. The test queries were divided so that each assessor judged, in total, roughly an equal number of documents. For each document, judges were asked to mark the document as "highly relevant", "somewhat relevant" or "not relevant" with respect to that query (as per TREC MedTrack guidelines). In addition, using Relevation!, assessors could provide a free-text comment regarding their decision. On completion of judging all documents for a query, the assessor was also asked to answer the following questions about the query: 1) "How difficult was this query to judge?". Choices: "Very difficult", "Moderately difficult" or "Easy". 2) "How would you rate the quality of the assessments you have provided for this query?" Choices: "High quality", "Average in quality" or "Poor quality". 3) "Other comments?" Here judges could provide qualitative comments regarding the particular query.

[1] 4 queries were excluded from the original 85 TREC MedTrack queries as no relevance assessments were collected for these.

As Relevation! is a web-based system, the HTTP access log was used to capture the interaction assessors had with the system. This included which queries and documents they viewed, when documents were judged and, importantly, the timestamps for these events. These timestamps were used to extract the amount of time each assessor spent in judging individual documents.[2] The difference in time between two consecutive HTTP POSTs was used as the measure of time it took to judge that document. On manual review, any time periods greater than 2500 seconds (42 minutes) was indicated as a break (e.g., lunch or coffee) and these timings were excluded. Note that qualitative feedback from assessors (e.g., difficulty and quality) were collected at query level, while quantitative statistics such as time to perform a judgement were collected both at query and document level.

A total of 58 hours (14.5 hours per assessor) of judging was required to complete the 942 documents.[3] The average time spent per document was 3.7 minutes. Using the control queries, inter-coder agreement was found to be 0.85, in-line with an inter-coder agreement of 0.8 found by the TREC MedTrack organisers.[4] Control queries also contained documents already judged by TREC assessors; therefore, if the TREC assessor is added as a fifth assessor, then agreement between all five assessors was 0.80.

3. IS ASSESSING RELEVANCE DEMANDING?

To determine if and why assessing relevance is demanding we analysed: (i) qualitative feedbacks given by assessors in relation to the assessment difficulty of each query; and (ii) the amount of time required to judge documents.

3.1 Did assessors find judging difficult?

Assessors rated each query according to how difficult it was to judge and further provided a self-assessment of the quality of their judgements. Results are shown in Figure 1. Assessors stated that about half of the queries were easy to assess, with the remaining half being of moderate difficulty. Only one query was considered very difficult to judge.[5] Nevertheless, the assessors believed the judgements they provided were of average or high quality. (No queries were marked as low quality.)

While these qualitative assessments are ultimately subjective (the self-perception of difficulty and quality may vary between assessors), it is clear how a significant number of assessments was perceived to be more demanding than others.

3.2 Time as indicator of demand

Beside examining the qualitative feedback of the difficulty in assessing documents, we also consider time as an indicator of judging demand. The intuition is that documents that required more time for assessment are more demanding; similarly, the longer it took on average to judge documents for a query, the more demanding that query.

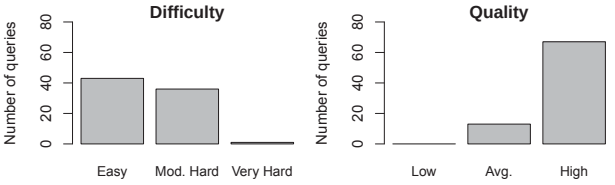

Figure 1: Judges' qualitative feedback on difficulty and quality of their assessments.

Difficulty	#Queries	Median sec./doc
Easy	44	130sec
Moderately Difficult	36	207sec (+59%)
Very Hard	1	219sec (+68%)

Table 1: Timing results by difficulty.

The use of time as an indicator of assessment demand is confirmed by the results of Table 1 that shows the judges' qualitative feedback about query difficulty along with the median document judging time for each difficulty level. This analysis shows that queries judged as moderately difficult took 59% longer to judge than those marked easy, endorsing the intuition that time is a (fine grain) indicator of assessment demand.

4. WHAT INDUCES COGNITIVE LOAD?

4.1 Are longer documents harder to judge?

Smucker & Clarke found that in web search, judging time was mainly influenced by document length [4]. Document length was, therefore, used as the main indicator for their time-biased evaluation measure [4].

In our study, if document length was also a measure of demand, then the Easy/Mod/Hard label assigned by assessors would simply relate to short, moderate and long documents respectively. By extension, shorter documents would be less demanding to judge. However, this was not found to be the case: there was no correlation between time to judge a document and the length of the document ($p = -0.0132$).

4.2 Are documents with discharge summaries easier to judge?

Many of the clinical documents used in our collection contained a discharge summary section.[6] Assessors commented that they often skimmed the document looking for a discharge summary section to read first rather than reading the document from top to bottom. Sometimes the relevance of a document could be determined from reading the discharge summary alone.[7] Based on these comments, we formed the hypothesis that documents containing a discharge summary would be quicker and less demanding to judge. However, our results show the contrary: the median time to judge a document with a discharge summary was 184 sec., vs. 118 sec. for documents without a discharge summary.

[2]The HTTP log is available online at:
https://github.com/ielab/MedIR2014-RelanceAssessment
[3]This number excludes documents from the control queries and those which took more than 2500 seconds to judge (i.e., where the assessors was deemed to have taken a break).
[4]Based on personal communication with Bill Hersh, TREC MedTrack organiser, 29 May 2013.
[5]Query 149: "Patients with delirium hypertension and tachycardia".
[6]A discharge summary is a narrative produced when a patient is discharged from hospital. Discharge summaries provide an overview of the patient's entire stay in hospital.
[7]Note that not all documents contained discharge summaries.

Documents	Time to judge (seconds)			
	mean	stddev	max	min
non-relevant	219	191	1614	5
relevant	224	221	2092	26
highly-relevant	167	209	2092	26
somewhat-relevant	289	217	1314	60

Table 2: Timing results by relevance grade.

4.3 Is the grade of relevance related to cognitive load?

Does the relevance grade of a document (i.e. highly relevant, relevant, not relevant) affect how demanding it is to judge? Table 2 shows the time it takes to judge documents according to the relevance grade. When considering only binary relevance (i.e., relevant vs. non-relevant), the average time to judge relevant and non-relevant documents does not differ significantly, although the time to judge relevant documents varies more (stddev); both the maximum and minimum judging time are greater for relevant documents. In contrast, when graded relevance is considered, some important differences are revealed: highly relevant documents are the *least* demanding to judge, whereas somewhat-relevant documents are the *most* demanding to judge. This finding suggests that clear cases of relevance (highly relevant or non-relevant) are less demanding. What is demanding is judging documents where relevance is less certain: cases where relevance is subjective or where the evidence for relevance is implicit and needs to be inferred. We explore more of these situations in the following section by analysing the assessors qualitative feedback.

5. WHY IS ASSESSMENT DEMANDING?

On completion of judging a query, assessors could optionally provide free-text, qualitative comments regarding their judging of the particular query. Assessors provided these comments for 57 out of 81 (70%) queries. We analysed their comments to gain a greater insight into their rational for assessment and to determine why it might be demanding. Table 3 contains a selection of assessor's comments which will be referred to throughout this section. The assessors' comments were used to identify queries exhibiting the following characteristics: (i) "objective", where the indicator of relevance was clear and explicit according to the assessor; (ii) "temporal", where relevance was strongly dependent on temporal aspects (of query or documents); (iii) "interpretation", where the interpretation of the query was subjective and the assessor had to decide on a particular interpretation; and (iv) "dependent aspects", where there were two or more conditions specified in the query — often dependent on each other — that had to be met. Queries not exhibiting any of the aforementioned characteristics were characteristed as "none". Note, these characteristics were derived from the relevance criteria, as stated in the assessor's comments, and not according to the query keywords. Queries were grouped according to these characteristics and we analysed the average time to judge the documents for queries with that characteristic. This is done to understand if some characteristics — and therefore some queries — were more demanding than others. The average time to judge according to each characteristic is shown in Figure 2.

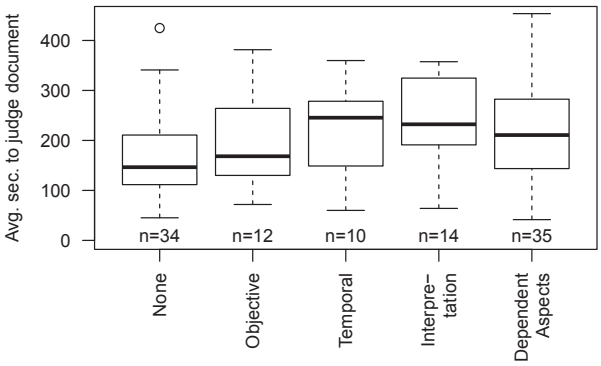

Figure 2: Average time to judge the documents for queries with different characteristics. Queries requiring some "interpretation" on the part of assessors were the most demanding.

Those queries identified as "none" (n=34, 60%) required, on average, the least assessment time and were the least demanding. Queries identified as "objective" (n=12, 21%) were marginally more demanding, as the assessor had a clear criteria to identify relevance and all that was required was to assert if that criteria applied to the particular document.

5.1 The effect of temporality on relevance

For "temporal" queries (n=10, 18%), the assessors specifically cited temporality as an important factor in determining relevance. The most common situation was when information pertaining to the query was found in the patient's past medical history section. Assessors had to decide whether the information was still valid: some conditions are ongoing (e.g., query `162`, Table 3), while others are temporal and are unlikely to still be valid (e.g., query `127`). In certain cases, assessors consulted the actual dates of the past medical history information to determine how recent the information was and whether it might still apply. In other cases, the query was interpreted according to a temporal definition (e.g., query `111`, where the assessor defined 'chronic back pain' as a condition persisting for at least 3 months). Queries exhibiting temporality tended to be the most demanding as assessors had to locate and reason with dates found in the documents.

5.2 Judging was highly subjective

For "interpretation" queries, assessors, at times, discussed their decisions regarding relevance. Although confident in their assessments, they stated that the interpretation of the query was subjective and often required careful consideration regarding different possible interpretations. For example, for query `101`, assessors debated whether a patient born deaf could be considered as exhibiting hearing loss. (Technically, if they never had any hearing, then they never had a loss of hearing.) One assessor thought such a document was relevant, while another assessor thought the document was not relevant. A medical encyclopaedia was consulted and the assessor decided to include patients born deaf as relevant. Queries requiring subjective interpretation showed a higher level of demand compared to other queries.

The task description given to assessors (recruitment of patients matching a certain inclusion criteria for clinical trials [5]) also affected their decisions regarding relevance. Certain documents described patients who had hearing loss

Query		Assessors' Comment
101	Patients with hearing loss	*It was not clear whether you wanted someone with current hearing loss or someone who had experienced reversible hearing loss due to an infection.*
102	Patients with complicated GERD who receive endoscopy	*Complicated GERD is a rather ambiguous term - could use clarification to yield better results (ex. stage a/b/c). Endoscopy is a blanket term for visualisation of a hollow organ - therefore some search results included patients who have had colonoscopies, but not upper endoscopies relevant to GERD.*
103	Hospitalized patients treated for methicillin resistant Staphylococcus aureus MRSA endocarditis	*Treatment of MRSA is the same no matter where it is in the body. Could have picked up a lot of documents because of the treatment regime or MRSA.*
111	Patients with chronic back pain who receive an intraspinal pain medicine pump	*The definition of chronic back pain used for these judgements was "greater than 3 months"*
127	Patients admitted with morbid obesity and secondary diseases of diabetes and or hypertension	*Without dates, it was difficult to ascertain whether or not hypertension and diabetes were secondary to patients' obesity, as is suggested by the query.*
162	Patients with hypertension on antihypertensive medication	*Once diagnosed with hypertension, you are generally considered to have it for the rest of your life …*
171	Patients with thyrotoxicosis treated with beta blockers	*A lot of hits for beta blockers and very few for any thyroid dysfunction.*
182	Patients with Ischemic Vascular Disease	*Straightforward to look at past medical history for coronary artery disease, bypass grafts or stents.*

Table 3: Assessors' qualitative comments regarding their experience judging the particular query.

on admission but the hearing loss was treated and resolved by discharge. In this case, assessors decided these patients would not be eligible for the clinical trial and, therefore, not relevant to the query. For other tasks (for example, finding how hearing loss is treated) these documents may have been highly relevant. These cases highlight the complex and often subjective nature of information need in this domain and that there are often implicit factors in the information need that do not transpire in the query. This further adds to the demand of relevance assessment for these types of queries.

5.3 Queries with dependent aspects

Queries with multiple "dependent aspects" received more debate by assessors and were also among the most demanding and those with the highest variance in judging time. The high variance in time to judge a document is due to the fact that queries with dependent aspects were either: (i) simple to judge, because the assessor just had to ascertain that a document met all aspects; or (ii) demanding to judge, because the assessor had to determine the interaction between the required aspects. Query 171 is an example of the former, simple case. Query 102 is an example of the latter case: GERD[8] is a common condition and is therefore found in many patients' records. The difficulty in interpreting this query was whether the endoscopy was performed because of the GERD or for some other, unrelated condition. There were a number of documents where patients had GERD but received the endoscopy for another reason; these were marked as not relevant. A similar query was 103, where endocarditis and MRSA were mentioned in the same document, but the cause of the endocarditis was not the MRSA. Again, these documents were marked as not relevant. These queries all have multiple dependent aspects to the query; even if both aspects are present in a document, that document may still not be relevant unless the dependence between them can be determined. Determining the dependence often required the assessors to exhaustively search through the document to identify the relationships between the dependent aspects. Doing so required longer judging times and was, therefore, more demanding.

[8]Gastroesophageal reflux disease (GERD) is caused when stomach acid comes up from the stomach into the esophagus.

6. CONCLUSION

Assessing relevance in medical IR is sometimes cognitively demanding and that demand differs depending on queries. Contrary to intuition and previous studies in other domains [4], this study found that document length does not influence demand. On the other hand, the grade of relevance is related with cognitive load (somewhat relevant documents were the most demanding to judge). Characteristics of queries that did increase demand included: temporality, subjectiveness of interpretation and the presence of multiple dependent aspects in the query.

A by-product of this study on what makes a relevance decision demanding, is the identification of some of the aspects that influence a relevance decision (for example, the role of temporality). Future work would, therefore, consider the actual features of the document (for example, temporal ranges or chronic vs. acute conditions) that identify these different aspects affecting relevance.

Data used in this study, including the HTTP interaction log, assessors' comments and qrels, is provided at: http://github.com/ielab/MedIR2014-RelanceAssessment.

Acknowledgements. The authors are grateful to Peter Bruza for his continued mentorship. The relevance assessments were conducted by Timothy Sladden, Warren Brown, Digvijay Khangarot and Thomas Souchen, from the University of Queensland.

7. REFERENCES

[1] Bevan Koopman. *Semantic Search as Inference: Applications in Health Informatics*. PhD thesis, Queensland University of Technology, 2014.

[2] B. Koopman and G. Zuccon. Relevation!: An open source system for information retrieval relevance assessment. In *SIGIR Demo*, Gold Coast, Australia, July 2014.

[3] S. Mizzaro. Relevance: The whole history. *JASIST*, 48(9):810–832, 1997.

[4] M. D. Smucker and C. L. Clarke. Time-based calibration of effectiveness measures. In *Proc. of SIGIR*, pages 95–104, Portland, U.S.A, 2012.

[5] E. M. Voorhees and W. Hersh. Overview of the trec 2012 medical records track. In *Proc. of TREC*, 2012.

Multi–modal relevance feedback for medical image retrieval

Dimitrios Markonis
HES–SO
TechnoPole 3
Sierre, Switzerland
dimitrios.markonis@hevs.ch

Roger Schaer
HES–SO
TechnoPole 3
Sierre, Switzerland
roger.schaer@hevs.ch

Henning Müller
HES–SO
TechnoPole 3
Sierre, Switzerland
henning.mueller@hevs.ch

ABSTRACT

Medical image retrieval can assist physicians in finding information supporting their diagnosis. Systems that allow searching for medical images need to provide tools for quick and easy navigation and query refinement as the time for information search is often short.

Relevance feedback is a powerful tool in information retrieval. This study evaluates relevance feedback techniques with regard to the content they use. A novel relevance feedback technique that uses both text and visual information of the results is proposed.

Results show the potential of relevance feedback techniques in medical image retrieval and the superiority of the proposed algorithm over commonly used approaches.

Future steps include integrating semantics into relevance feedback techniques to benefit of the structured knowledge of ontologies and experimenting on the fusion of text and visual information.

Keywords

relevance feedback, content–based image retrieval, medical image retrieval

1. INTRODUCTION

Searching for images is a daily task for many medical professionals, especially in image–oriented fields such as radiology. However, the huge amount of visual data in hospitals and the medical literature is not always easily accessible and physicians have generally little time for information search as they are charged with many tasks.

Therefore, medical image retrieval systems need to return information adjusted to the knowledge level and expertise of the user in a quick and precise fashion. A well known technique trying to improve search results by user interaction is relevance feedback [13]. Relevance feedback allows the user to mark results returned in a previous search step as relevant or irrelevant to refine the initial query. The concept behind relevance feedback is that though user may have difficulties in formulating a precise query for a specific task, they generally see quickly whether a returned result is relevant to the information need or not. This technique found use in image retrieval particularly with the emerge of content–based image retrieval (CBIR) systems [18, 19, 20]. Following the CBIR mentality, the visual content of the marked results is used to refine the initial image query. With the result images represented as a grid of thumbnails, relevance feedback can be applied quickly to speed up the search iterations and refine results. Recent user–tests with radiologists on a medical image search system also showed that this method is intuitive and straightforward to learn [7].

Depending on whether the user manually provides the feedback to the system (e.g. by marking results) or the system obtains this information automatically (e.g. by log analysis) relevance feedback can be categorized as explicit or implicit. Moreover, the information obtained by relevance feedback can be used to affect the general behaviour of the system (long–term learning). In [11] a market basket analysis algorithm is applied in image retrieval of long–term learning. A recent review of short–term and long–term learning relevance feedback techniques in CBIR can be found in [6]. An extensive survey of relevance feedback in text–based retrieval systems is presented in [15] and for CBIR in [14].

In the medical informatics field, [1] applies CBIR with relevance feedback on mammography retrieval. In [12], an image retrieval framework using relevance feedback is evaluated on a dataset of 5000 medical images that uses support vector machines to compute the refined queries.

In this paper we evaluate different explicit, short–term relevance feedback techniques using visual content or text for medical image retrieval. We propose a technique that combines visual and text–based relevance feedback and show that it achieves a competitive performance to the state–of–the–art approaches.

2. METHODS

2.1 Rocchio algorithm

One of the most well known relevance feedback techniques is Rocchio's algorithm [13]. Its mathematical definition is given below:

$$\vec{q}_m = \alpha\vec{q}_o + \beta\frac{1}{|D_r|}\sum_{\vec{d}_j \in D_r}\vec{d}_j - \gamma\frac{1}{|D_{nr}|}\sum_{\vec{d}_j \in D_{nr}}\vec{d}_j \quad (1)$$

where $\vec{q}_m$ is the modified query,
$\vec{q}_o$ is the original query,
D_r is the set of relevant images,

MedIR 2014, July 11, 2014, Gold Coast, Australia.

D_{nr} is the set of non–relevant images and
α, β and γ are weights.

Typical values for the weights are $\alpha = 1, \beta = 0.8$ and $\gamma = 0.2$. Rocchio's algorithm is typically used in vector models and also for CBIR. Intuitively, the original query vector is moved towards the relevant vectors and away from the irrelevant ones. By giving a weight to the positive and negative parts a problem of CBIR can be avoided that when more negative than positive feedback exists that also many relevant images disappear from the results set.

2.2 Late fusion

Another technique that showed potential in image retrieval [5] is late fusion. Late fusion [2] is used in information retrieval to combine result lists. It can be applied for fusing multiple features, multiple queries and in multi–modal techniques. The concept behind this method is to merge the result lists into a single list while boosting common occurrences using a fusion rule.

For example, the fusion rule of the score–based late fusion method CombMNZ [17] is defined as:

$$S_{\texttt{combMNZ}}(i) = F(i) * S_{\texttt{combSUM}}(i) \quad (2)$$

where $F(i)$ is the number of times an image i is present in retrieved lists with a non–zero score, and $S(i)$ is the score assigned to image i. CombSUM is given by

$$S_{\texttt{combSUM}}(i) = \sum_{j=1}^{N_j} S_j(i) \quad (3)$$

where $S_j(i)$ is the score assigned to image i in retrieved list j.

2.3 Multi–modal relevance feedback

Most of the techniques use vectors either from the text or the visual models. However, it has been shown that approaches that use both text and visual information can outperform single–modal ones in image retrieval. We propose the use of multi–modal information for relevance feedback to enhance the retrieval performance. This is, to the extend of our knowledge, the first time that such a technique is proposed in image retrieval. As late fusion is applied on result lists, it is straightforward to use for combining results from visual and text queries.

2.4 Experimental setup

For evaluating the relevance feedback techniques the following experimental setup was followed: The n search iterations are initiated with a text query in iteration 0. The relevant results from the top k results of iteration i were used in the relevance feedback formula of the iteration $i+1$ for $i = 0...n-2$.

The image dataset, topics and ground truth of ImageCLEF 2012 medical image retrieval task [9] were used in this evaluation. The dataset contains more than 300'000 images from the medical open access literature.

The image captions were accessed by the text–based runs and indexed with the Lucene[1] text search engine. Vector space model was used along with tokenization, stopword removal, stemming and Inverse document frequency-Term frequency weighting. The Bag–of–visual–words model described in [3] and the bag–of–colors model appearing in [4]

[1]http://lucene.apache.org/

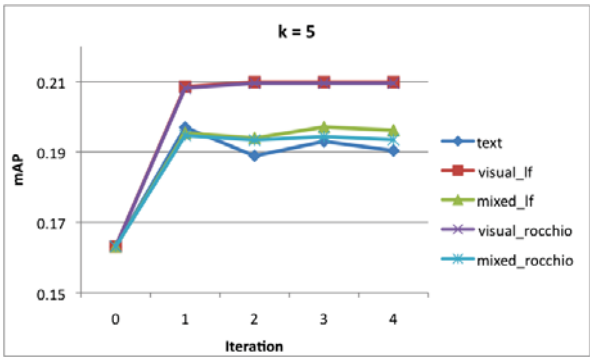

Figure 1: Mean average precision per search iteration for $k = 5$.

Table 1: Best mAP scores

Run	k = 5	k = 20	k = 50	k = 100
text	0.197 (1)	0.2544 (4)	0.3107 (3)	0.3349 (4)
visual_lf	0.2099 (2)	0.2243 (3)	0.2405 (4)	0.2553 (3)
visual_roc	0.2096 (2)	0.2187 (2)	0.2249 (3)	0.2268 (2)
mixed_lf	0.1971 (3)	0.2606 (4)	0.3079 (4)	0.3487 (3)
mixed_roc	0.1947 (1)	0.2635 (4)	0.3207 (4)	0.3466 (4)

were used for the visual modelling of the images. In multi-modal runs, the fusion of the visual and text information is performed only for the text 1000 top results as in the evaluation of ImageCLEF only the top 1000 documents are taken into account in any case.

Five techniques were evaluated in this study:

1. **text**: text–based RF using vector space model. Word stemming, tokenization and stopword removal is performed in both text and multi–modal runs.
2. **visual_rocchio**: visual RF using Rocchio to fuse the relevant image vectors and CombMNZ fusion to fuse the original query's results with the visual ones.
3. **visual_lf**: visual RF using late fusion (and the CombMNZ fusion rule) to fuse the relevant image results and the original query results with the visual ones.
4. **mixed_rocchio**: multimodal RF using Rocchio to fuse the relevant image vectors and CombMNZ fusion to fuse the original query results with the relevant caption results and relevant visual results.
5. **mixed_lf**: multimodal RF using late fusion (and the CombMNZ fusion rule) to fuse the relevant image results and the original query results with the captions' results and relevant visual results.

3. RESULTS

The evaluation of the five techniques was performed for $k = 5, 20, 50, 100$ and $n = 5$. Results of the mean average precision (mAP) of each technique per iteration are shown in Figures 1, 2, 3, 4.

Table 1 gives the best mAP scores of each run. The numbers in parentheses are the number of the iteration when this score was achieved. For scores that were the same in multiple iterations of the same run, the iteration closer to the first is used.

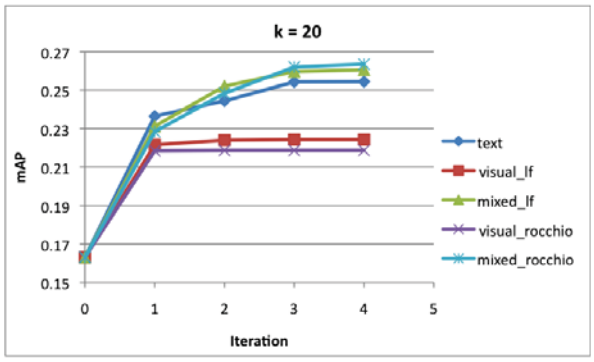

Figure 2: Mean average precision per search iteration for $k = 20$.

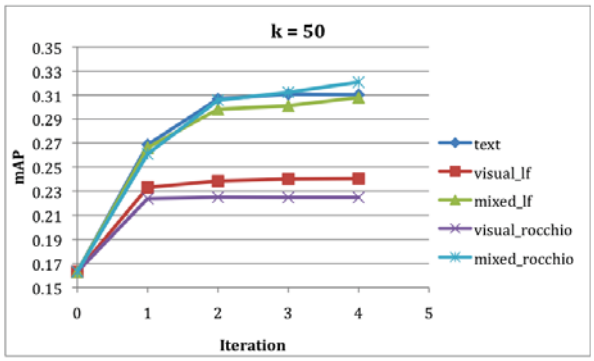

Figure 3: Mean average precision per search iteration for $k = 50$.

4. DISCUSSION

All of the evaluated techniques improve retrieval after the initial search iteration. This demonstrates the potential of relevance feedback for refining medical image search queries.

Relevance feedback using only visual appearance models, even though improving the retrieval performance after the first iteration, performed worse than the text–based runs in most cases. Visual features still suffer from the semantic gap between the expressiveness of visual features and our human interpretation. Still, this shows their usefulness in image datasets where no or little text meta–data are available. Moreover, when combined with the text–information in the proposed method, they improve the text–only baseline.

The proposed multi–modal runs provide the best results in all the cases except for case $k = 5$. Surprisingly, the visual runs perform slightly better than the text and the multi–modal approaches for this case. However, assuming independent and normal distributed average precision values the significance tests show that the difference is not statistically significant.

We consider the case $k = 20$ as the most realistic scenario since users do not often inspect more than 2 pages of results. Especially for grid–like result interface views, where each page can contain 20 to 50 results, we consider $k = 20$ more realistic than $k = 5$. In this case the proposed methods achieve the best performance with 0.2606 and 0.2635 respectively. Again, the significance tests do not find any

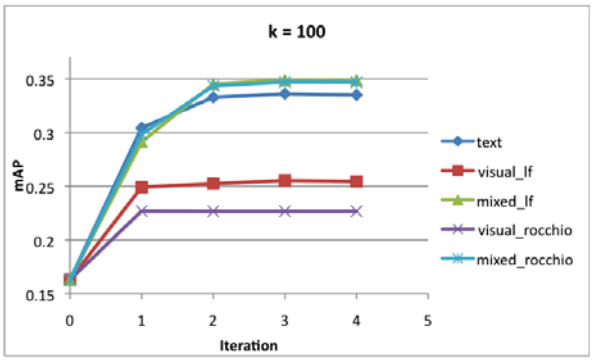

Figure 4: Mean average precision per search iteration for $k = 100$.

significance difference between the three best approaches. However, applying different fusion rules for combining visual and text information (such as linear–weighting) could further improve the results of the mixed approaches.

It can be noted that as the k increases, the performance improvement also increases, highlighting the added value of relevance feedback. Larger values of k were not explored as this scenarios were judged as unrealistic.

In the visual runs using Rocchio for combining the visual queries is performing worse than late fusion. This comes in accordance with the findings in [3]. The reason behind this could be that the large visual diversity of relevant images in medicine and the curse of dimensionality cause the modified vector to behave as an outlier in the high dimensional visual feature space. In the mixed runs the difference between the two methods is not statistically significant with Rocchio performing slightly better than the late fusion.

Irrelevant results were ignored, as they often have little or no impact on the retrieval performance [10, 16]. More importantly, the ground truth of the dataset used contains a much larger portion of annotated irrelevant results than relevant ones. This was considered to potentially simulate an unrealistic scenario, as users do not usually mark many results as negative examples. Having too many negative examples could also cause the modified vector to follow an outlier behaviour. Preliminary results confirmed this hypothesis, where the use of negative results for relevance feedback can decrease performance after the first iteration.

It should be noted that this is an automated relevance feedback experiment of positive only feedback and that in selective relevance feedback situations the retrieval performance is expected to perform even better. A larger number of steps could be investigated but this might be unrealistic, given the fact that physicians have little time and stop after a few minutes of search [8]. Often users will only test a few steps of relevance feedback at the most.

5. CONCLUSIONS

This paper proposes the use of multi–modal information when applying relevance feedback to medical image retrieval. An experiment was set up to simulate the relevance feedback of a user on a number of medicine–related topics from ImageCLEF 2012.

In general, all the techniques evaluated in this study improve the performance, which shows the added value of rele-

vance feedback. Text–based relevance feedback showed consistently good results. Visual–based techniques showed competitive performance for small shortlist sizes, underperforming in the rest of the cases. The proposed multi–modal approaches showed promising results slightly outperforming the text–based one but without statistical significance.

More fusion techniques are going to be evaluated in the future. Comparison to manual query refinement by users is considered in future plans, to assess relevance feedback as a concept in medical image retrieval. The addition of semantic search is also of interest, to take advantage of the structured knowledge of the medical ontologies such as RadLex (Radiology Lexicon) and MeSH (Medical Subject Headings).

6. ACKNOWLEDGEMENTS

This work was supported by the EU 7th Framework Program in the context of the Khresmoi project (grant 257528).

7. REFERENCES

[1] C.-C. Chen, P.-J. Huang, C.-Y. Gwo, Y. Li, and C.-H. Wei. Mammogram retrieval: Image selection strategy of relevance feedback for locating similar lesions. *International Journal of Digital Library Systems (IJDLS)*, 2(4):45–53, 2011.

[2] A. Depeursinge and H. Müller. Fusion techniques for combining textual and visual information retrieval. In H. Müller, P. Clough, T. Deselaers, and B. Caputo, editors, *ImageCLEF*, volume 32 of *The Springer International Series On Information Retrieval*, pages 95–114. Springer Berlin Heidelberg, 2010.

[3] A. García Seco de Herrera, D. Markonis, I. Eggel, and H. Müller. The medGIFT group in ImageCLEFmed 2012. In *Working Notes of CLEF 2012*, 2012.

[4] A. García Seco de Herrera, D. Markonis, and H. Müller. Bag of colors for biomedical document image classification. In H. Greenspan and H. Müller, editors, *Medical Content–based Retrieval for Clinical Decision Support*, MCBR–CDS 2012, pages 110–121. Lecture Notes in Computer Sciences (LNCS), Oct. 2013.

[5] A. García Seco de Herrera, D. Markonis, R. Schaer, I. Eggel, and H. Müller. The medGIFT group in ImageCLEFmed 2013. In *Working Notes of CLEF 2013 (Cross Language Evaluation Forum)*, September 2013.

[6] J. Li and N. M. Allinson. Relevance feedback in content-based image retrieval: a survey. In *Handbook on Neural Information Processing*, pages 433–469. Springer, 2013.

[7] D. Markonis, F. Baroz, R. L. Ruiz de Castaneda, C. Boyer, and H. Müller. User tests for assessing a medical image retrieval system: A pilot study. In *MEDINFO 2013*, 2013.

[8] D. Markonis, M. Holzer, S. Dungs, A. Vargas, G. Langs, S. Kriewel, and H. Müller. A survey on visual information search behavior and requirements of radiologists. *Methods of Information in Medicine*, 51(6):539–548, 2012.

[9] H. Müller, A. García Seco de Herrera, J. Kalpathy-Cramer, D. Demner Fushman, S. Antani, and I. Eggel. Overview of the ImageCLEF 2012 medical image retrieval and classification tasks. In *Working Notes of CLEF 2012 (Cross Language Evaluation Forum)*, September 2012.

[10] H. Müller, W. Müller, D. M. Squire, S. Marchand-Maillet, and T. Pun. Strategies for positive and negative relevance feedback in image retrieval. Technical Report 00.01, Computer Vision Group, Computing Centre, University of Geneva, rue G n ral Dufour, 24, CH–1211 Gen ve, Switzerland, Jan. 2000.

[11] H. Müller, D. M. Squire, and T. Pun. Learning from user behavior in image retrieval: Application of the market basket analysis. *International Journal of Computer Vision*, 56(1–2):65–77, 2004. (Special Issue on Content–Based Image Retrieval).

[12] M. M. Rahman, P. Bhattacharya, and B. C. Desai. A framework for medical image retrieval using machine learning and statistical similarity matching techniques with relevance feedback. *Information Technology in Biomedicine, IEEE Transactions on*, 11(1):58–69, 2007.

[13] J. J. Rocchio. Relevance feedback in information retrieval. In *The SMART Retrieval System, Experiments in Automatic Document Processing*, pages 313–323. Prentice Hall, Englewood Cliffs, New Jersey, USA, 1971.

[14] Y. Rui, T. S. Huang, and S. Mehrotra. Relevance feedback techniques in interactive content–based image retrieval. In I. K. Sethi and R. C. Jain, editors, *Storage and Retrieval for Image and Video Databases VI*, volume 3312 of *SPIEProc*, pages 25–36, Dec. 1997.

[15] I. Ruthven and M. Lalmas. A survey on the use of relevance feedback for information access systems. *The Knowledge Engineering Review*, 18(02):95–145, 2003.

[16] G. Salton and C. Buckley. Improving retrieval performance by relevance feedback. *Readings in information retrieval*, 24:5, 1997.

[17] J. A. Shaw and E. A. Fox. Combination of multiple searches. In *TREC-2: The Second Text REtrieval Conference*, pages 243–252, 1994.

[18] D. M. Squire, W. Müller, H. Müller, and T. Pun. Content–based query of image databases: inspirations from text retrieval. *Pattern Recognition Letters (Selected Papers from The 11th Scandinavian Conference on Image Analysis SCIA '99)*, 21(13–14):1193–1198, 2000. B.K. Ersboll, P. Johansen, Eds.

[19] L. Taycher, M. L. Cascia, and S. Sclaroff. Image digestion and relevance feedback in the ImageRover WWW search engine. pages 85–94, 1997.

[20] M. E. Wood, N. W. Campbell, and B. T. Thomas. Iterative refinement by relevance feedback in content–based digital image retrieval. pages 13–20, 1998.

A Joint Local-Global Approach for Medical Terminology Assignment

Liqiang Nie
National University of Singapore
nieliqiang@gmail.com

Mohammad Akbari
National University of Singapore
akbari@nus.edu.sg

Tao Li
Zhejiang University
coylee917@gmail.com

Tat-Seng Chua
National University of Singapore
chuats@nus.edu.sg

ABSTRACT

In community-based health services, vocabulary gap between health seekers and community generated knowledge has hindered data access. To bridge this gap, this paper presents a scheme to label question answer(QA) pairs by jointly utilizing local mining and global learning approaches. Local mining attempts to label individual QA pair by independently extracting medical concepts from the QA pair itself and mapping them to authenticated terminologies. However, it may suffer from information loss and lower precision, which are caused by the absence of key medical concepts and presence of irrelevant medical concepts. Global learning, on the other hand, works towards enhancing the local mining via collaboratively discovering missing key terminologies and keeping off the irrelevant terminologies by analyzing the social neighbors. Practically, this unsupervised scheme holds potential to large-scale data.

Categories and Subject Descriptors

J.3 [**Life and Medical Sciences**]: Health

Keywords

Community-based Health Services, Question Answers, Vocabulary Gap, Medical Terminology Assignment

1. BACKGROUND

The rise of digital technologies has transformed the patient-doctor relationships. Nowadays, when patients struggle with their health concerns, the majority usually explore the Internet to research the problem before and after they see their doctors. For example, 70% of Canadians turned to Internet to look up health-related information in *2009* [8] and 72% of American Internet users searched for health solutions in *2012* [4]. These metrics have reflected the scope and scale of the online health seekers.

Figure 1: The illustration of a QA example from community-based health services (HealthTap).

To better serve the needs of health seekers, community-based health services have emerged as effective platforms for health knowledge dissemination and exchange, such as HealthTap[1], HaoDF[2] and WenZher[11]. They not only permit health seekers to freely post health-oriented questions, but also encourage doctors to provide trustworthy answers. Figure 1 demonstrates one typical QA pair example. Over time, a tremendous number of QA pairs has been accumulated in their repositories, and in most circumstances, health seekers may directly locate good answers by searching from these archives, rather than waiting for the experts' responses or painfully browsing through a list of documents from the general search engines.

2. CHALLENGES

In many cases, the community generated health content may not be directly usable due to the vocabulary gap, since participants with diverse backgrounds do not necessarily share the same vocabulary. Take HealthTap as an example. The same question may be described in substantially different ways by two individual health seekers. On the other hand, the answers provided by doctors may contain acronyms with multiple possible meanings, and non-standardized terms.

[1]https://www.healthtap.com/
[2]www.haodf.com

MedIR 2014, July 11, 2014, Gold Coast, Australia.

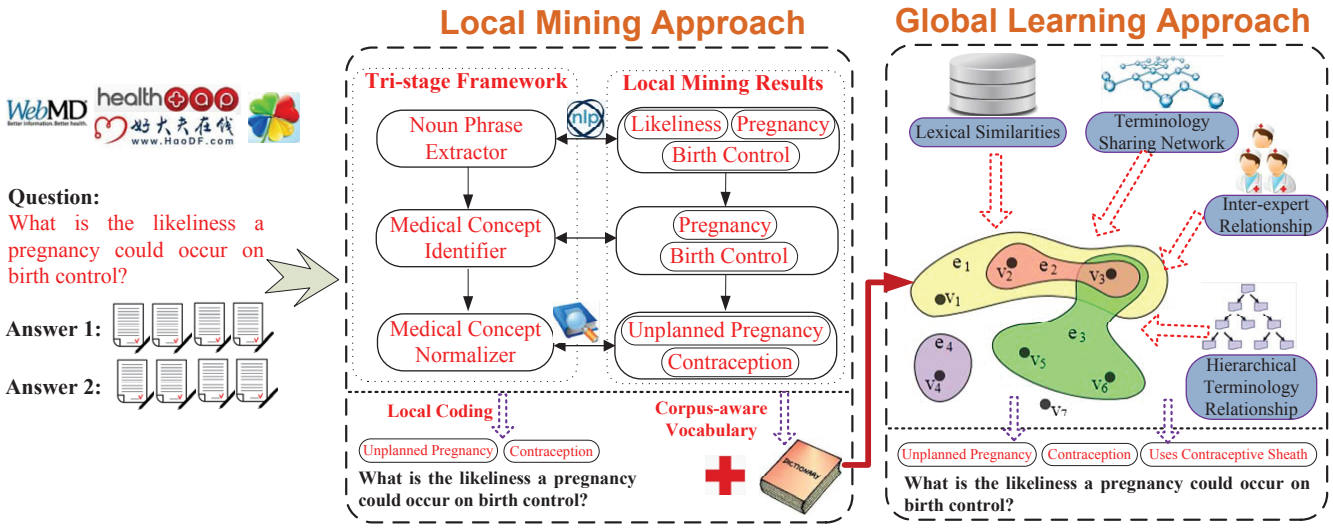

Figure 2: The schematic illustration of the proposed automatic medical terminology assignment scheme. The answer part is not displayed due to the space limitation.

In this work, we define medical concepts as medical domain-specific noun phrases, and medical terminologies as authenticated phrases by well-known organizations that are used to accurately describe the human body and associated components, conditions and processes in a science-based manner. Even though some health communities have recently suggested doctors to annotate their answers with medical concepts, we cannot ensure that they are medical terminologies. Meanwhile, the tags adopted by doctors often vary greatly [3]. For example, "heart attack" and "myocardial disorder" are employed by different doctors to refer to the same medical diagnosis. It was shown that the inconsistency of community generated health data greatly hindered the cross-resource data exchange, management and integrity [9]. Even worse, it was reported that users had encountered big challenges in reusing the archived content due to the incompatibility between their search terms and those accumulated medical records [21]. Therefore, automatic coding of the QA pairs with standardized terminologies is highly desired. It leads to a consistent interoperable way of indexing, storing and aggregating across specialties and sites. In addition, it facilitates QA pair retrieval via bridging the vocabulary gap between the queries and archives by coding the new queries with the standardized terminologies.

It is worth mentioning that there already exist several efforts dedicated to research on automatically mapping medical records to terminologies [19, 2, 10, 7, 17]. Most of these efforts, however, focused on hospital generated health data or health provider released sources by utilizing either isolated or loosely coupled rule-based and machine learning approaches. Compared to this kind of data, the emerging community generated health data is more colloquial, in terms of inconsistency, complexity and ambiguity, which pose challenges for data access and analytics. Further, most of the previous work simply utilizes the external medical dictionary to code the medical records rather than considering the corpus-aware terminologies. Their reliance on the external corpus independent knowledge may potentially bring in inappropriate terminologies. Constructing a corpus-aware terminology vocabulary to prune the irrelevant terminologies of specific dataset and narrow down the candidates is the tough issue we are facing. In addition, the varieties of heterogeneous cues were often not adequately exploited simultaneously. Therefore, a robust integrated framework to draw the strengths from various resources and models is still expected.

3. METHOD

To overcome these limitations, we propose a novel scheme that is able to code the QA pairs with corpus-aware terminologies. As illustrated in Figure 2, the proposed scheme consists of two mutually reinforced components, namely, local mining and global learning.

3.1 Local Mining

Local mining aims to locally code the QA pairs by extracting the medical concepts from individual instance and then mapping them to terminologies based on the external authenticated vocabularies. To accomplish this task, we establish a tri-stage framework, which includes noun phrase extraction, medical concept detection and medical concept normalization.

To extract all the noun phrases, we initially assign part-of-speech tags to each word in the given QA pair by Stanford POS tagger[3]. We then extract tag sequences that match a fixed pattern of part-of-speech tags as noun phrases from the texts. This pattern is formulated as follows.

$$(Adjective|Noun)^*(Noun \quad Preposition) \\ ?(Adjective|Noun)^*Noun. \tag{1}$$

A sequence of tags matching this pattern ensures that the corresponding words make up a noun phrase. For example, the following complex sequence can be extracted as a noun phrase: "ineffective treatment of terminal lung cancer".

Inspired by the efforts in [18, 6], in order to differentiate the medical concepts from other general noun phrases, we assume that concepts that are relevant to medical domain occur frequently in medical domain and rarely in

[3]http://nlp.stanford.edu/software/tagger.shtml

non-medical ones. Based on this assumption, we employ the concept entropy impurity (CEI) [6] to comparatively measure the domain-relevance of a concept by comparing the term frequencies between two different corpora D_1 and D_2. D_1 is our medical-domain corpus and D_2 is a general English Gigaword data of Linguistic Data Consortium[4].

As aforementioned, we cannot ensure that all medical concepts are standardized terminologies. Take "birth control" as an example. It is recognized as a medical concept by our approach, but it is not an authenticated terminology. Instead, we should map it into "contraception". Therefore, it is essential to normalize the detected medical concepts according to an appropriate external standardized dictionary and this normalization is the key to bridging the vocabulary gap. In this work, we use SNOMED CT[5] as our dictionary, since it provides the core general terminologies for the electronic health record and formal logic-based hierarchical structure. The terminologies and their descriptions in SNOMED CT are first indexed[6]. We then search each medical concept against the indexed SNOMED CT. For the medical concepts with multiple matched results, e.g., two results returned for "female", we keep all the returned terminology candidates for further selection. Enlightened by Google distance [1], we estimate the semantic similarity between the medical concept and the returned terminology candidates via exploring their co-occurrence on Google. We then select the most relevant terminology candidate as the normalized result.

Local mining, however, may suffer from various problems. The first problem is incompleteness. This is because some key medical concepts may not explicitly present in the QA pairs. The QA pair illustrated in Figure 2 shows an example of this situation, where the accurate terminology: "use contraceptive sheath" is absent from the QA pair. The second one is the lower precision. This is due to some irrelevant medical concepts explicitly embedded in the QA pairs, and are mistakenly detected and normalized by the local approach. For instance, given the question, "*What are the risks getting pregnant and giving birth later in life ?*", the terminology "finding of life event" as normalized from the irrelevant medical concept "life" is assigned as code. It is less informative to capture the main intent.

3.2 Global Learning

It is noteworthy that most previous efforts, including our local approach, attempted to map the QA pairs directly to the entries in external dictionaries without any pruning. This approach often presents problems since the external dictionaries usually cover relatively comprehensive terminologies and are far beyond the vocabulary scope of the given corpus. It may result in the deterioration in coding performance in terms of efficiency and effectiveness. The problem is caused by the over-widened scope of vocabularies, which may bring in unpredictable noises and make the precise terminology selection challenging. As a byproduct, a corpus-aware terminology vocabulary is naturally constructed by our local mining approach, which can be used as terminology space for further learning.

Let $\mathcal{Q} = \{q_1, q_2, ..., q_N\}$ and $\mathcal{T} = \{t_1, t_2, ..., t_M\}$ respectively denote a repository of QA pairs and their associated locally mined terminologies. The target of global learning is to learn appropriate terminologies from the global terminology space $\mathcal{T}$ to annotate each q in $\mathcal{Q}$. In this work, the global learning task is regarded as a multi-label learning problem[16]. It is formulated as,

$$\arg\min_{\mathbf{F}} \sum_{i=1}^{M} \left\{ \Omega(\mathbf{f}_i) + \lambda L(\mathbf{f}_i) + \mu \sum_{j=1}^{M} R_{ij} \|\mathbf{f}_i - \mathbf{f}_j\|^2 \right\}, \quad (2)$$

where M refers to the number of classes, i.e., the number of medical terminologies to be assigned. Vector $\mathbf{f}_i$ is the ith column of $\mathbf{F}$, representing the relevance scores of each QA pair to the i-th terminology. $\Omega(\mathbf{f})$ and $L(\mathbf{f})$ denotes the regularizer on the hypergraph and empirical loss, respectively. In addition, R_{ij} is the inter-terminology relationship between terminology i and terminology j. They are mined by exploiting the external well-structured ontology, which are able to alleviate the granularity mismatch problems and reduce the irrelevant sibling terminologies. By differentiating the above equation with respect to $\mathbf{F}$, we can obtain a closed-form solution.

The philosophy to formulate these three objectives is as follows. The first objective aims to guarantee that the relevance probability function is continuous and smooth in semantic space. This means that the relevance probabilities of semantically similar QA pairs should be close to each other. The second objective is ensured by the empirical loss function, which forces the relevance probabilities to approach the initial roughly estimated relevance scores. These two implicit constraints are widely adopted in reranking-oriented approaches [12, 13, 14, 15]. The last encourages the values of QA pairs, which are connected by hierarchical structured terminologies, to be similar to each other.

When it comes to hypergraph construction, the N QA pairs from $\mathcal{Q}$ are regarded as vertices and they are connected by three types of hyperedges. The first type takes each vertex as a centroid and forms a hyperedge by circling around its k-nearest neighbors based on QA pair content similarities. This procedure was first adopted in [5]. The second type is based on terminology-sharing network. For each terminology, it groups all the QA pairs sharing the same terminology together. The third type actually takes the users' social behaviours into consideration by rounding up all the questions answered by closely associated doctors. The inter-doctor relationships are inferred from the doctors' historical data. Specifically, doctors who are frequently respond to the same kinds of questions probably share highly overlapping expertise, and thus the questions they answered can be regarded as semantically similar to a certain extent. As a consequence, up to $N + M + U$ hyperedges are constructed in our hypergraph, where U is the number of involved doctors. Learning from this hypergraph, we are able to find missing key concepts and propagate precise terminologies among underlying connected records over a large collection. Besides the semantic similarity among QA pairs and terminology-sharing network, the inter-terminology and inter-expert relationships are seamlessly integrated in the proposed model. It is noteworthy that a rich set of healthcare specific features are extracted and weighted for similarity estimation.

4. EXPERIMENTS

We crawled more than 109 thousand QA pairs from

[4] http://www.ldc.upenn.edu/
[5] http://www.ihtsdo.org/snomed-ct/
[6] http://viw2.vetmed.vt.edu/sct/menu.cfm

Table 1: The comparative evaluation results of medical terminology assignment in terms of $S@K$ and $P@K$.

Metric / Apporach	S@1	S@2	S@3	S@4	P@1	P@2	P@3	P@4
LocalMining	72.0%	84.0%	91.0%	95.0%	72.0%	72.1%	69.7%	68.3%
Local+Global	**83.0%**	**92.0%**	**98.0%**	**100.0%**	**83.0%**	**81.5%**	**80.3%**	**78.8%**

Table 2: Comparative illustration of the representative question samples with locally mined terminologies and locally+globally recommended terminologies. Answers are not displayed due to limited space.

QA pairs	Locally Mined Terminologies	Local Mining + Global Learning
Is it safe to color my hair during pregnancy ?	hair structure, dyed hair, feeling safe, patient currently pregnant, first trimester pregnancy...	hair structure, patient currently pregnant, coal tar allergy, hair color change, disorder of endocrine system...
If I get an infection caused by gum disease, can that be transferred to my fetus ?	infectious disease, gingival disease, entire fetus, inflammation, periodontal disease...	infectious disease, prematurity of fetus, gingival disease, periodontal disease low birth weight infant...

HealthTap, which involve 5,958 unique doctors. For ground truth construction, we invited three professionals with master degrees majored in medicine programme. The labelers were trained with a short tutorial and a set of demonstrating examples. A majority voting scheme among the three labelers can partially alleviate the subjectivity problem. The annotators were required to label only top five recommended terminologies for each QA pair, and they were labeled either as "positive" or "negative". 100 QA pairs were labeled as testing set.

We adopted two metrics that are able to characterize precisions from different aspects. The first is average $S@K$ over all testing QA pairs, which measures the probability of finding a relevant terminology among the top K recommended ones. To be specific, for each testing QA pair, $S@K$ is assigned to 1 if a relevant terminology is positioned in the top K and 0 otherwise. The second one is average $P@K$ that stands for the proportion of recommended terminologies that are relevant[20]. $P@K$ is defined as $P@K = \frac{|\mathcal{C} \cap \mathcal{R}|}{|\mathcal{C}|}$

where $\mathcal{C}$ is a set of the top K terminologies, and $\mathcal{R}$ is the manually labeled positive ones.

Table 1 displays the comparison. We can see that the local mining approach achieves the worst performance. This is reasonable, because irrelevant concepts may be mapped to terminologies because of their presence in the QA pairs.

Table 2 comparatively illustrates the representative QA pair samples with locally minded terminologies and locally+globally recommended ones. Intuitively, the terminologies are more comprehensive and reliable after enhancement with global learning.

5. CONCLUSIONS AND FUTURE WORK

This paper presented a medical terminology assignment scheme to bridge the vocabulary gap between health seekers and community generated knowledge. A strong unified framework of local mining and global learning is proposed to tackle this research issue, instead of the conventional isolated utilization. It proposes the concept entropy impurity approach to comparatively detect and normalize the medical concepts locally, which naturally construct a corpus-aware terminology vocabulary with the help of external knowledge. In addition, it builds a novel global learning model to enhance the local coding results. This model seamlessly integrates various heterogeneous cues.

In the future, we will investigate how to flexibly organize the unstructured medical content into user needs-aware ontology by the recommended medical terminologies.

6. ACKNOWLEDGEMENTS

This work was supported by NUS-Tsinghua Extreme Search project under the grant number: R-252-300-001-490.

7. REFERENCES

[1] R. L. Cilibrasi and P. M. B. Vitanyi. The google similarity distance. *IEEE Transactions on Knowledge and Data Engineering*, 2007.

[2] C. Dozier, R. Kondadadi, K. Al-Kofahi, M. Chaudhary, and X. Guo. Fast tagging of medical terms in legal text. In *Proceedings of the International Conference on Artificial Intelligence and Law*, 2007.

[3] A. e-HIM Work Group on Computer-Assisted Coding. Delving into computer-assisted coding. *Journal of American Health Information Management Association*, 2004.

[4] S. Fox and M. Duggan. Health online 2013. Survey, Pew Research Center, 2013.

[5] Y. Huang, Q. Liu, S. Zhang, and D. Metaxas. Image retrieval via probabilistic hypergraph ranking. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2010.

[6] M.-Y. Kim and R. Goebel. Detection and normalization of medical terms using domain-specific term frequency and adaptive ranking. In *Information Technology and Applications in Biomedicine, IEEE International Conference on*, 2010.

[7] L. S. Larkey and W. B. Croft. Automatic assignment of icd9 codes to discharge summaries. *PhD Thesis, University of Massachusetts at Amherst*, 1995.

[8] M. Law. Online drug information in canada. Technical report, 2012.

[9] G. Leroy and H. Chen. Meeting medical terminology needs-the ontology-enhanced medical concept mapper. *IEEE Transactions on Information Technology in Biomedicine*, 2001.

[10] L. V. Lita, S. Yu, S. Niculescu, and J. Bi. Large scale diagnostic code classification for medical patient records. In *Proceedings of the Conference on Artificial Intelligence in Medicine*, 1995.

[11] L. Nie, T. Li, M. Akbari, J. Shen, and T.-S. Chua. Wenzher: Comprehensive vertical search for healthcare domain. In *Proceedings of the International ACM SIGIR Conference*, 2014.

[12] L. Nie, M. Wang, Y. Gao, Z.-J. Zha, and T.-S. Chua. Beyond text qa: Multimedia answer generation by harvesting web information. *IEEE Transactions on Multimedia*, 2013.

[13] L. Nie, M. Wang, Z.-J. Zha, and T.-S. Chua. Oracle in image search: A content-based approach to performance prediction. *ACM Transactions on Information System*, 2012.

[14] L. Nie, M. Wang, Z.-J. Zha, G. Li, and T.-S. Chua. Multimedia answering: Enriching text qa with media information. In *Proceedings of the International ACM SIGIR Conference*, 2011.

[15] L. Nie, S. Yan, M. Wang, R. Hong, and T.-S. Chua. Harvesting visual concepts for image search with complex queries. In *Proceedings of the International Conference on Multimedia*, 2012.

[16] L. Nie, Y.-L. Zhao, X. Wang, J. Shen, and T.-S. Chua. Learning to recommend descriptive tags for questions in social forums. *ACM Transactions on Information System*, 2014.

[17] H. Suominen, F. Ginter, S. Pyysalo, A. Airola, T. Pahikkala, S. Salanterä, and T. Salakoski. Machine learning to automate the assignment of diagnosis codes to free-text radiology reports: a method description. In *Proceedings of the ICML Workshop on Machine Learning for Health-Care Applications*, 2008.

[18] P. Velardi, M. Missikoff, and R. Basili. Identification of relevant terms to support the construction of domain ontologies. In *Proceedings of the workshop on Human Language Technology and Knowledge Management*, 2001.

[19] L. Yves A., S. Lyudmila, and F. Carol. Automating icd-9-cm encoding using medical language processing: A feasibility study. In *Proceedings of the AMIA Annual Symposium*, 2000.

[20] Y.-L. Zhao, L. Nie, X. Wang, and T.-S. Chua. Personalized recommendations of locally interesting venues to tourists via cross region community matching. *ACM Transactions on Intelligent Systems and Technology*, 2013.

[21] G. Zuccon, B. Koopman, A. Nguyen, D. Vickers, and L. Butt. Exploiting medical hierarchies for concept-based information retrieval. In *Proceedings of the Seventeenth Australasian Document Computing Symposium*, 2012.

Exploring Clustering Based Knowledge Discovery towards Improved Medical Diagnosis

Rajendra Prasath
Dept. of Business Information Systems, University College Cork
Ireland
R.Prasath@ucc.ie

Philip O'Reilly
Dept. of Business Information Systems, University College Cork, Ireland
Philip.OReilly@ucc.ie

ABSTRACT

We propose to develop a framework for an intelligent reasoner with capabilities that support complex decision making processes in medical diagnosis. *Identifying* the causes, *reasoning* the effects to explore information geometry and *learning* the associated factors, from medical forum information extracted, are the core aspects of this work. As part of the proposed framework, we present an approach that identifies semantically similar causes and effects for any specific disease from medical diagnosis literature using implicit semantic interconnections among the medical terms. First we crawled MedHelp[1] forum data and considered two types of information: *forums* data and *posts* data. Each forum link points to a specific disease and consists of several topics pertaining to that disease. Each topic consists of multiple posts that carry either users' queries/difficulties or doctor's feedback pertaining to the issue(s) of the users. We use graph based exploration on the terms (diseases) and their relations (in terms of causes/effects) and explore the information geometry pertaining to similar diseases. We performed a systematic evaluation to identify the relevance of the contextual information retrieved for a specific disease and similar factors across different diseases. The proposed approach looks promising in capturing similar causes and/or effects that pertain to multiple diseases. This would enable medical practitioners to have a multi-faceted view of a specific disease/condition.

Keywords

Causes and Effects, Medical Diagnosis, Semantically Similar diseases, Information Geometry, Graph Analysis

1. INTRODUCTION

Understanding the causes and effects pertaining to a specific disease is key to better prediction in medical diagnosis and improved patient management. Diseases/Conditions may have similar or semantically related causes and effects. Furthermore, gaining insight on how diseases are diagnosed and managed would enable medical practitioners to make more informed decisions on disease management. Illustrating the role of machine learning as an enabler of this for more informed decision support is a key aspect of this paper. In this paper, we present an approach to identify causes and effects that exist across different diseases using a graph clustering based knowledge discovery approach.

Understanding causation and correlation between individual factors is key to decision making in multiple domains including medicine and business. Traditionally, such association between elements has been identified through human interpretation of text. However, this is a very time consuming, manual, labour intense process and is limited by human capacity. The ability to mine textual content, identify association and the nature of that association between elements using machine learning techniques provides significant opportunities. Specifically in the medical domain, where a significant amount of content is textual in nature (e.g. medical notes), having the ability to identify causation and correlation between elements in large medical datasets provides significant opportunity for advancing medical research and enabling better decision making pertaining to condition diagnosis and patient management.

In this paper, we attempt to identify and create term clusters using graph clustering approach and then perform topic classification. This approach improves the document classification task by putting the terms, that are semantically related, in the same cluster. Users could search for a specific disease and explore information pertaining to its causes and effects by means of semantically related texts.

2. CLUSTERING BASED KNOWLEDGE DISCOVERY

To incorporate natural language understanding, commonsense and domain specific knowledge could be used to improve the text representation by including more generated informative features to perform deep understanding of the document text than the mere *Bag-of-Words* approach [4, 2, 8]. Mitra *et* al. [6] proposed an unsupervised feature selection algorithm suitable for data sets, based on measuring similarity between features whereby redundancy therein is removed. Pedersen and Kulkarni [7] presented a system called SenseClusters to cluster similar contexts in natural language text and assigns identifying labels to these clusters based on their content. In addition to clustering similar contexts, it can be used to identify

[1] http://www.medhelp.org/forums/list

MedIR 2014, July 11, 2014, Gold Coast, Australia.
ACM SIGIR.

synonyms and sets of related words[2]. To incorporate this kind of additional common-sense/domain knowledge, Gabrilovich and Markovitch [3] used world knowledge from open source knowledge repository like Wikipedia to generate additional features. Similar intuition is adopted to form word clusters that generate features enriching the document content in a better way. Then the documents are represented in the knowledge-rich space of generated features. This leads to better organization of semantically related text representation. In this proposed scheme, given a knowledge repository, the text documents are examined and their representation is enriched in a completely mechanical way. Motivated by the above considerations, our aim is to empower machine learning techniques for text representation with a substantially wider body of knowledge like the one obtained from the superior inference capabilities of humans.

2.1 Mathematical Formulation

In this section, we characterize the medical forum data in a formal way. Each post pertaining to a specific topic is informally written and we focus on terms and their co-occurrences. We have n textual descriptions, viz-a-viz, posts: $P = \{p_1, p_2, \cdots, p_n\}$ and each post can be formally represented as a sequence of terms, illustrating the scenario of the underlying disease, as follows: $p_i = \{t_1, t_2, \cdots, t_m\}$ where $1 \leq i \leq m$ and m varies differently for different post. We convert the entire text data of all posts into a graph, say $G = (V, E)$ where V represents the set of nodes (each term in P is considered as a node) and the co-occurrence of any pair of nodes across the posts is considered to be the edge representing the strength of association between the pair of nodes.

2.2 Graph Clustering

Clustering deals with identifying a pattern/structure in the bunch of unlabeled data. In general, clustering organizes data into groups whose members are related in some way and two or more data can be grouped into the same cluster if they are, in some way, falling close to each others' context. Clustering has many useful applications like finding a group of people with similar behavior, processing orders, grouping plants and animals, grouping web blog data to access similar patterns.

While exploring a variety of possibilities to identify the context of causes and effects of diseases from text fragments, feature space grows and it is hardly possible to limit the expansion of the new feature space containing the local contexts extracted from the informal writing of medical data. This could possibly be solved by using dimensionality reduction techniques to limit the size of the document to be classified. This attempt first makes word cluster vectors using unsupervised feature generation by identifying the related contexts. Then using the identified contexts, supervised learning is performed for categorizing the given text collection consisting of user posts.

First, we use the entire collection of medical diagnosis related posts and filter out the list of the distinguishable unique terms. Using these terms, we first build the weighted graph in which nodes represent terms and edges represent the weight - the number of documents in which the given pair of terms co-occurs across the collection of posts. For each unique term, the list of documents in which it occurs is retrieved. Using this data, we build the weighted graph in which the edge between two terms would represent their semantic association implicitly. This process is repeated for all features and a weighted graph for the overall data is constructed. Thus the problem is modeled into a graph clustering problem. This results in a graph $G = (V, E, A)$ where $|V| = n$ represents the number of unique terms; $|E|$ represents the number of edges and the adjacency matrix; and A is $|V| \times |V|$ whose nonzero entries correspond to the edge weight between a pair of terms (adjacency list is assumed in case of sparse matrix - in this case, number of rows in the graph represents the total number of terms in the graph).

We use the kernel-based multilevel clustering algorithm proposed by Dhillon *et* al. [1] on the weighted input graph with the number of desired partitions. This algorithm uses three steps: coarsening, base-clustering, and refinement. In coarsening, the given graph is repeatedly transformed into smaller subgraphs. This process is repeated until a few nodes remain in the graph. Then during base clustering, regional growing approach [5] could be used with these few nodes. The quality of the resulting clusters depends on the choice of the initial nodes. The refinement process is applied as follows: If a node in G_i is in cluster c, then all nodes in G_{i-1} formed from that node are in cluster c. For more details, please refer to [1, 5].

In this work, we generate term clusters from extracted word graphs, using co-occurrence information of terms. The task is to partition the graph into clusters so that terms could be grouped into a few subsets and dimensionality of the new term space is reduced. Now based on the generated term clusters, we perform classification to identify similar causes and effects across various diseases.

2.3 Proposed Approach

The proposed approach works as follows: From the posts of each topic, textual descriptions are extracted. Unique terms (after removing stop words) are considered as nodes in the graph and the number of times a pair of terms co-occurs in the entire corpus is considered as the weight of the edge connecting the pair of terms. At first, we build word cluster vectors using graph clustering algorithm.

Algorithm 1 Building Word Clusters

Input: A set of n textual descriptions (posts)
$P = \{p_1, p_2, \cdots, p_n\}$
A set of predefined category labels $C = \{c_1, c_2, \cdots, c_l\}$

Build Word Cluster Vectors:

1: Extract text from posts and build the unique word list
2: **for** each unique term t_i in the word list **do**
3: Identify the existence of edges from t_i to all other terms with nonzero positive weight.
4: Store the co-occurring term with its corresponding edge weight in the adjacency list
5: **end for**
6: Use kernel-based multilevel graph clustering algorithm on the adjacency list and perform clustering to generate cluster IDs
7: For every cluster ID, construct word clusters
8: Store these word cluster vectors

[2] *word* and *term* are used interchangibly

Secondly, we use these word cluster vectors to re-represent text documents that discuss the causes and effects of a specific disease. Then we perform classification of the re-represented text documents. The proposed algorithm inherently applies clustering semantically related terms and performs classification of them. Based on the word clusters found in the first step, we perform the classification on the clustered space of label pertaining to the terms in the original documents.

Algorithm 2 Classification using Word Clusters

1: Preprocess the text documents by removing numbers, punctuations and stop-words (using SMART[3] word list)

2: **for** each processed text (post) data p_i in P **do**
3: **for** each unique term in p_i **do**
4: Identify its cluster id
5: Map the given feature in terms of its cluster ID
6: Augment text fragments with cluster ID mappings
7: **end for**
8: **end for**
9: Build classifier on these mapped/expanded text data (containing only cluster IDs) and use it to predict the class of the text having similar causes and effects
10: Compute the classification accuracy

Output: The category label(s) for texts (post) that have similar causes and effects across diseases.

3. EXPERIMENTAL RESULTS

3.1 Corpus

We crawled a subset of MedHelp[4] forum data from the world wide web. The MedHelp forum is organized as a set of topics, each representing a specific disease and under each topic, there are several subtopics. Each subtopic consists of several posts from both users (may be patients or their dependents) and doctors as well.

For our experiments, we have selected 15 categories covering the most widely discussed topics in the MedHelp forum. The details of this experimental corpus is given in Table. 1

We have used 3 different types of classifiers, namely *Naive Bayes*, *k−Nearest Neighbours (k−NN)*, and *Support Vector Machines (SVM)*, to test the effect of the proposed approach. We have used *rainbow*[5] to build the classification models and during classification, we have used 60% of the data for training and 40% for testing.

3.1.1 Evaluation Methodology

We have used Precision, Recall, F-Measure and the classification Accuracy to evaluate the quality of the identified diseases having similar causes and effects. We use the two-way confusion matrix given in Table. 2 to derive the evaluation measures:

Precision is defined as follows:

$$Precision(P) = \frac{TP}{TP + FP}$$

[4] http://www.medhelp.org/forums/list/

[5] The 'Bow' Toolkit - http://www.cs.cmu.edu/~mccallum/bow/

S.No	Class	#Diseases (Posts)
1	Asthma	35 (43)
2	Breast-Cancer	35 (37)
3	COPD	41 (47)
4	Cosmetic	57 (64)
5	Dental	97 (100)
6	Embarazo	41 (52)
7	Genetic-Disorder	59 (68)
8	Hepatitis	110 (147)
9	Kidney	69 (89)
10	Liver-Transplant	64 (62)
11	Oral	35 (47)
12	Pathology	31 (36)
13	Respiratory	48 (54)
14	Thyroid-Cancer	58 (63)
15	Varicose-Veins	36 (52)

Table 1: Corpus Statistics

	Correctly Classified	Wrongly Classified
Actual=Yes	True Positive (TP)	False Negative (FN)
Actual=No	False Positive (FP)	True Negative (TN)

Table 2: 2-way confusion matrix

Recall is measured by the following equation:

$$Recall(R) = \frac{TP}{TP + FN}$$

F-Measure is computed by considering the ratio between Precision and Recall and hence calculated as follows:

$$F_\beta = \frac{(\beta^2 + 1)PR}{\beta^2 P + R}$$

The balanced F_1-measure with $\beta = 1$ (when P and R are weighted equally), is given by

$$F_1 = \frac{2pr}{P + R}$$

Classification accuracy is measured from confusion table as follows:

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

3.1.2 Discussion

We have observed that the proposed approach performs well in classifying the medical text having causes and issues expressed by either the patients or their friends or relatives. Figure. 1 shows the classification accuracy of proposed approach with top 15 classes using three different types of classifiers. While using the Naive Bayes method, the accuracy for the class "Asthma" goes down due to the fact that certain specific causes are pertaining to the "Respiratory" related disease. Similar misclassification takes place across "Dental" and "Oral" classes and these misclassified instances share common causes. Even though, the diseases like "Breast-cancer" and "Thyroid-cancer" share a common cause for cancer, misclassification is significantly less. Additionally the misclassification takes place across the diseases: "Hepatitis" and "Liver-Transplant".

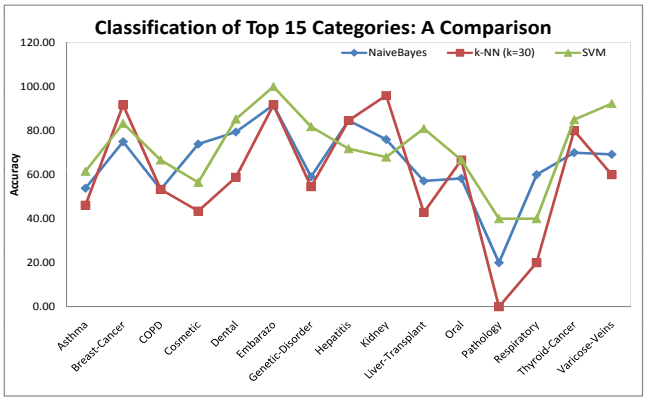

Figure 1: Classification of top 15 classes: A comparison of Naive Bayes, $k-$NN and SVM methods

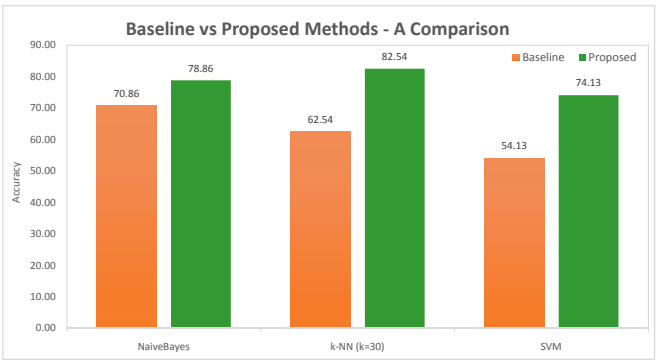

Figure 2: Comparison of the overall classification accuracy: Baseline vs Proposed approaches

We have observed the classification accuracy of $k-$NN classifier for various values of k (=10, 20, 30, 50) and found that for k =30, the system performs very well. In this classification task, we have observed that the maximum number of instances from the class "Liver-Transplant" is misclassified under "Hetatitis" class as the causes of these common diseases coincide by and large. At the same time, none of the instances is classified correctly for the class "Pathology" as these causes and effects are very similarly described as that of "Hepatitis" and "Thyroid-cancer". While using the SVM classifier, we noticed that some instances are misclassified across the classes: "Breast-cancer" and "Cosmetic". In this case, user raised cross reference related queries with post surgical treatment of the Breast-cancer disease. Similar misclassification is found across the classes: "Chronic Obstructive Pulmonary Disease" (COPD) and "Resporatory". Subsequently, we would like to apply this approach for effective retrieval of causes and effects pertaining to a specific disease. Also we will draw the information geometry of prominent diseases that share the common causes/effects in our subsequent experiments.

4. CONCLUSION

We proposed a method to enable greater understanding of various conditions, their symptoms, treatment and management by *identifying* similar scenarios, *reasoning* the effects to explore information geometry and *learning* the associated contextual factors, from medical forum information extracted from health services data. This approach identifies semantically similar causes and effects for any specific disease/condition, using implicit semantic interconnections among the medical terms. We use graph based exploration on the terms and their relations (causes/effects) across the collection of posts and explore the information geometry pertaining to the similar diseases. We evaluated the relevance of the contextual information retrieved for a specific disease and/or similar factors across different diseases. The proposed approach looks promising in capturing similar scenarios pertaining to multiple diseases. This would enable medical practitioners to have a multi-faceted view about any specific disease, towards better decision making.

Acknowledgments: This research is co-funded by the Irish Government (Enterprise Ireland) and European Union (European Regional Development Fund). Dr. Philip O'Reilly is the Principal Investigator responsible for this research and can be contacted at `Philip.Oreilly@ucc.ie`.

5. REFERENCES

[1] Dhillon, I. S., Guan, Y., and Kulis, B. Weighted graph cuts without eigenvectors a multilevel approach. *IEEE Trans. Pattern Anal. Mach. Intell. 29*, 11 (2007), 1944–1957.

[2] Gabrilovich, E. *Feature Generation for Textual Information Retrieval Using World Knowledge*. PhD thesis, Technion - Israel Institute of Technology, Haifa, Israel, 2006.

[3] Gabrilovich, E., and Markovitch, S. Feature generation for text categorization using world knowledge. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence* (San Francisco, CA, USA, 2005), IJCAI'05, Morgan Kaufmann Publishers Inc., pp. 1048–1053.

[4] Giles, J. Internet encyclopaedias go head to head. *Nature 438*, 1 (2005), 900–901.

[5] Karypis, G., and Kumar, V. A fast and high quality multilevel scheme for partitioning irregular graphs. *SIAM J. Sci. Comput. 20*, 1 (1998), 359–392.

[6] Mitra, P., Murthy, C. A., and Pal, S. K. Unsupervised feature selection using feature similarity. *IEEE Trans. Pattern Anal. Mach. Intell. 24*, 3 (2002), 301–312.

[7] Pedersen, T., and Kulkarni, A. Identifying similar words and contexts in natural language with senseclusters. In *Proc. of the 20th national conf. on Artificial intelligence* (2005), AAAI'05, AAAI Press, pp. 1694–1695.

[8] Prasath, R., and Sarkar, S. Unsupervised feature generation using knowledge repositories for effective text categorization. In *Proceedings of the 2010 Conference on ECAI 2010: 19th European Conference on Artificial Intelligence* (Amsterdam, The Netherlands, The Netherlands, 2010), IOS Press, pp. 1101–1102.

Integrating Understandability in the Evaluation of Consumer Health Search Engines

Guido Zuccon
Queensland University of Technology
Brisbane, Australia
g.zuccon@qut.edu.au

Bevan Koopman
Australian e-Health Research Centre, CSIRO
Brisbane, Australia
bevan.koopman@csiro.au

ABSTRACT

In this paper we propose a method that integrates the notion of understandability, as a factor of document relevance, into the evaluation of information retrieval systems for consumer health search. We consider the gain-discount evaluation framework (RBP, nDCG, ERR) and propose two understandability-based variants (uRBP) of rank biased precision, characterised by an estimation of understandability based on document readability and by different models of how readability influences user understanding of document content. The proposed uRBP measures are empirically contrasted to RBP by comparing system rankings obtained with each measure. The findings suggest that considering understandability along with topicality in the evaluation of information retrieval systems lead to different claims about systems effectiveness than considering topicality alone.

Categories and Subject Descriptors: H.3 [Information Storage and Retrieval]: H.3.3 Information Search and Retrieval
General Terms: Evaluation.

1. INTRODUCTION

Searching for health advice on the Web is an increasingly common practice. A recent research has found that in 2012 about 58% of US adults (72% of all US Internet users – 66% in 2011) have consulted the Internet for health advice [5]; of these, 77% have used search engines like Google, Bing, or Yahoo! to gather health information, while only 13% have started their health information seeking activities from specialised sites such as WebMD. It is, therefore, crucial to create and evaluate information retrieval (IR) systems that specifically support consumers searching for health advise on the Web. In this paper, we focus on the evaluation of IR systems for consumer health search.

Previous studies within health informatics have investigated online consumer health information beyond topicality to specific health topics; in particular, with respect to the understandability and reliability of such information. For example, Wiener and Wiener-Pla [11] have investigated the readability (measured by the SMOG reading index [7]) of Web pages concerning pregnancy and the periodontium as retrieved by Google, Bing and Yahoo!. Walsh and Volsko [10] have shown that most online information sampled from five US consumer health organisations and related to the top 5 medical related causes of death in US is presented at a readability level (measured by the SMOG, FOG and Flesch-Kincaid reading indexes [7]) that exceeds that of the average US citizen (7th grade level). Ahmed et al. [1] have highlighted the variability in readability (measured by the Flesch Reading Ease and the Flesch-Kincaid reading index [7]) and quality of concussion information accessed through Google searches. The understandability and reliability of online health information has been considered as a critical issue for supporting online consumer health search because (1) consumers may not benefit from health information that is not provided in an understandable way; and (2) the provision of unreliable, misleading or false information on a health topic, e.g., a medical condition or treatment, may led to negative health outcomes. This previous research suggests that topicality should not be considered as the only relevance factor for assessing the effectiveness of IR systems for consumer health search: other factors, such as understandability and reliability, should also be included in the evaluation framework.

Research on the user perception of document relevance has shown that users' relevance assessments are affected by a number of factors beyond topicality, although topicality has been found to be the essential relevance criteria. For example, Xu and Chen proposed and validated a five-factor model of relevance which consists of novelty, reliability, understandability, scope, along with topicality [12]. Their empirical findings highlight the importance of understandability, reliability and novelty along with topicality in the relevance judgements they collected. Nevertheless, typical evaluation of IR systems commonly considers only relevance assessments in terms of topicality[1]; this is also the case when evaluating systems for consumer health search, for example, within CLEF eHealth 2013 [6]. In this paper, we aim to close this gap in the evaluation of IR systems and focus on integrating understandability along with topicality for the evaluation of consumer health search engines. The integration of other factors influencing relevance, such as reliability, are left for future work.

The integration of understandability within the evaluation methodology is achieved by extending the general gain-discount framework synthesised by Carterette [3]; this framework encompasses the widely-used nDCG, RBP and ERR. The result is a series of understandability-biased evaluation measures. Specifically, we examine one such measure, the understandability-based rank biased precision (uRBP) – a variant of rank biased precision (RBP) [8]; variants of nDCG

[1]With the recent exception of novelty and diversity, e.g., [4].

MedIR July 11, 2014, Gold Coast, Australia.
ACM SIGIR.

and ERR may also be derived within our framework.

The proposed evaluation measure is further instantiated by considering specific estimations of understandability based on readability measures computed for each retrieved document. While understandability encompasses other aspects in addition to text readability (e.g., prior knowledge), the use of readability measures is a good first approximation for understandability. This choice is also supported by prior work in health informatics regarding understandability of consumer health information (e.g., see [11, 10, 1]).

The impact of the proposed framework and the specific resultant measures on the evaluation of IR systems is investigated in the context of the consumer health search task of CLEF eHealth 2013 [6]; empirical findings show that systems that are most effective according to uRBP are not necessarily as effective when considering topicality alone (i.e. RBP).

2. UNDERSTANDABILITY-BASED EVALUATION

2.1 The gain-discount framework

We tackle the problem of jointly evaluating topicality and understandability for measuring IR system effectiveness within the gain-discount framework synthesised by Carterette [3]. Within this framework, the effectiveness of a system, conveyed by a ranked list of documents, is measured by the evaluation measure M, defined as:

$$M = \frac{1}{\mathcal{N}} \sum_{k=1}^{K} g(k)d(k) \tag{1}$$

where $g(k)$ and $d(k)$ are respectively the gain and discount function computed for the document at rank k,[2] K is the depth of assessment at which the measure is evaluated, and $1/\mathcal{N}$ is a (optional) normalisation factor, which serves to bound the value of the sum into the range [0,1] (see [9]).

Different measures developed within the gain-discount framework are characterised by different instantiations of its components. For example, the discount function in RBP is modelled by $d(k) = \beta^{k-1}$, where $\beta \in [0, 1]$ reflects user behaviour (high values representing persistent users, low values representing impatient users); while in nDCG the discount function is given by $d(k) = 1/(\log_2(1 + k))$ and in ERR by $d(k) = 1/k$. Similarly, instantiations of gain functions differ depending upon the considered measure. In RBP, the gain function is binary-valued (i.e., $g(k) = 1$ if the document at rank k is relevant, $g(k) = 0$ otherwise); while for nDCG $g(k) = 2^{r(k)} - 1$ and for ERR $g(k) = (2^{r(k)} - 1)/2^{r_{max}}$ (with $r(k)$ being the relevance grade of the document at rank k).

Without loss of generality, we can express the gain provided by a document at rank k as a function of its probability of relevance; for simplicity we shall write $g(k) = f(P(R|k))$, where $P(R|k)$ is the probability of relevance given the document at rank k. Note that a similar form has been used for the definition of the gain function for time-biased evaluation measures [9]. The specific instantiations of $g(k)$ in measures like RBP, nDCG and ERR can be seen as the application of different functions $f(.)$ to estimations of $P(R|k)$.

Traditional TREC-style relevance assessors are instructed to consider topicality as the only (explicit) factor influencing relevance, thus $P(R|k) = P(T|k)$, i.e., the probability that the document at k is topically relevant (to a query).

[2]For simplicity of notation, in the following we override k to represent the *rank position* k, or the *document* at rank k: the context of use will determine the meaning of k.

2.2 Integrating understandability

As discussed by previous work, e.g. [12], relevance is influenced by many factors; topicality being only one of them – although the most important. To integrate understandability into the gain-discount framework, we model $P(R|k)$ as the joint $P(T, U|k)$, i.e. the probability of relevance of a document (at rank k) is estimated using the joint probability of the document being topical and understandable.

To compute the joint probability we assume that topicality and understandability are compositional events and their probabilities independent, i.e., $P(T, U|k) = P(T|k)P(U|k)$. This is a strong assumption and its limitations are briefly discussed in Section 4. Following this assumption, the gain function in the gain-discount framework is expressed as:

$$g(k) = f(P(R|k)) = f\big(P(T|k)P(U|k)\big) \tag{2}$$

Different evaluation measures that may be developed within this framework would instantiate $f\big(P(T|k)P(U|k)\big)$ in different ways. In the following we will propose two RBP-based instantiations; other instantiations are left for future work.

2.3 Estimating understandability

In the traditional TREC settings, assessments about the topicality of a document to a query are collected through manual annotation of query-document pairs from assessors (i.e., binary or graded relevance assessments[3]); these are then turned into estimations of $P(T|k)$. This process may be mimicked to collect understandability assessments; in this paper however we do not explore this possibility. Instead, we explore the possibility of computing understandability as a property of a document and integrate this in the evaluation process, along with standard relevance assessments. To this aim, readability is used as a proxy for understandability. (The limitations of this choice are briefly noted in Section 4.) Its use is however justifiable because readability is one of the aspects that influence the understanding of text.

To estimate readability (and thus understandability), we employ established general readability measures as those used in [1, 10, 11], e.g., SMOG, FOG and Flesch-Kincaid reading indexes. These measures consider the surface level of the text contained in Web pages, that is, wording and syntax of sentences. In this framework, the presence of long sentences, words containing many syllables and unpopular words, are all indicators of difficult text to read [7]. In this paper, we use the FOG measure to estimate the readability of a text; the FOG reading level is computed as

$$FOG(d) = 0.4 * (avgslen(d) + phw(d)) \tag{3}$$

where $avgslen(d)$ is the average length of sentences in a document d and $phw(d)$ is the percentage of hard words (i.e., words with more than two syllables) in d.

The use of such general readability measures to assess the readability of documents concerning health information has been questioned [13] as these do not seem to adequately correlate with human judgments for documents in this domain [13]. Nevertheless, the adoption of standard readabil-

[3]Recall that although called "relevance assessments", in TREC-style assessments, annotators are usually instructed to consider only the topicality of a document to a query, isolating this factor from others influencing relevance in real settings.

ity measures in this paper is a first step towards demonstrating the use of the proposed understandability biased measures and analyse how system rankings would change accordingly. In addition, their usage is partially supported by previous work within health informatics on assessing the readability of online health advice [1, 10, 11].

2.4 Modelling P(U|k)

Given the readability score for a document at rank k, $P(U|k)$ needs to be estimated; this is achieved by considering user models that encode different ways in which a user is affected by document readability.

We first consider a user model $P_1(U|k)$ where a user is characterised by a readability threshold th and every document that has a readability score below th is considered certainly understandable, i.e., $P_1(U|k) = 1$; while documents with readability above th are considered not understandable, i.e. $P_1(U|k) = 0$. This is a (Heaviside) step function centred in th; this function is depicted in Figure 1 ($P_1(U|k)$) with $th = 20$, along with the FOG readability score distribution for documents from CLEF e-Health 2013 [6]. The use of a step function to model $P(U|k)$ is akin to the gain function in RBP (also a step function). The understandability-based RBP for user model one is then given by:

$$uRBP_1 = (1-\beta)\sum_{k=1}^{K}\beta^{k-1}r(k)u_1(k) \tag{4}$$

where, for simplicity of notation, $u_1(k)$ indicates the value of $P_1(U|k)$ and $r(k)$ is the (topical) relevance assessment of document k (alternatively, the value of P(T|k)); thus $g(k) = f(P(T|k)P(U|k)) = P(T|k)P(U|k) = r(k)u_1(k)$.

A second user model ($P_2(U|k)$) is proposed, where the probability estimation is similar to a step function, but smoothed in the surroundings of the threshed value; this provides a more realistic transition between readable and not-readable content:

$$P_2(U|k) \propto \frac{1}{2} - \frac{\arctan\Big(FOG(k) - th\Big)}{\pi} \tag{5}$$

where arctan is the arctangent trigonometric function and $FOG(k)$ is the FOG readability score of document at rank k; other readability scores could be used instead of FOG. The distribution of $P_2(U|k)$ values is shown in Figure 1. Equation 5 is not a proper probability distribution, but this can be obtained by normalising Equation 5 by its integral between $[\min\big(FOG(k)\big), \max\big(FOG(k)\big)]$; however Equation 5 is rank equivalent to such distribution, not changing the effect on the uRBP variant. These settings lead to the formulation of a second understandability-based RBP, $uRBP_2$, based on the second user model, by simply substituting $u_2(k) = P_2(U|k)$ to $u_1(k)$ in Equation 4.

Note that in both understandability-based measures (as well as in the original RBP) the contribution of an irrelevant document is zero, irrespective of its $P(U|k)$. The contribution (to the gain) of a relevant document with readability score above th is 1 for RBP, 0 for $uRBP_1$ and less than 0.5 for $uRBP_2$ (for $uRBP_2$ the score will quickly tend to 0 the more the readability score is above the threshold value).

Finally, note that it is possible to design other user models representing how readability influences document understandability; the challenge is to determine which model better represents the relationship between readability and document understanding.

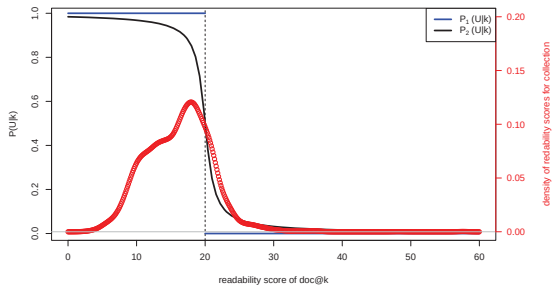

Figure 1: Distributions for $\mathbf{P_1(U|k)}$ and $\mathbf{P_2(U|k)}$ with respect to threshold $\mathbf{th = 20}$, along with the density distribution of readability scores (computed using FOG) for the documents in the CLEF eHealth 2013 qrels.

3. EMPIRICAL ANALYSIS

3.1 Experiment design and settings

To understand how accounting for understandability influences the evaluation of IR systems tailored to searching health advice on the Web, we consider the runs submitted to the CLEF eHealth 2013 [6], which specifically aimed at evaluating systems for this task. Our empirical experiments and subsequent analysis specifically focus on the changes in system rankings obtained when evaluating with standard measures (RBP) and understandability-based measures ($uRBP_1$ and $uRBP_2$). System rankings are compared using Kendall rank correlation (τ) and AP correlation [14] (τ_{AP}), which weights higher rank changes that affect top systems. We do not experiment with different values of β in RBP, and set $\beta = .95$ across RBP and uRBP.

The document collection used in CLEF eHealth 2013 has been retired due to removal of duplicates and copyrighted documents; we thus use the CLEF eHealth 2014 collection (which is a subset of the CLEF eHealth 2013 collection) to allow reproducibility of the reported results and the 2013 qrels for relevance assessment. For each document in the collection, the FOG readability scores (Equation 3) were computed – the score distribution for all documents in the CLEF eHealth 2013 qrels is shown in Figure 1. Three thresholds on the FOG readability values were explored for the computation of the two alternative formulations of uRBP: $th = 10, 15, 20$; documents with a FOG score below 10 should be near-universally understandable, while documents with FOG scores above 15 and 20 increasingly restrict the audience able to understand the text.

3.2 Results and analysis

Figure 2 reports RBP vs. uRBP of IR systems participating to CLEF eHealth 2013 for the two user models proposed in Section 2.4 and for the three readability thresholds considered in the experiments. Similarly, Table 1 reports the values of Kendall rank correlation (τ) and AP correlation (τ_{AP}) between system rankings obtained with RBP and the two versions of uRBP.

Higher correlation between systems rankings obtained with RBP and uRBP is observed for higher values of th, irrespectively of uRBP version (see Table 1). This is expected as the higher the threshold, the more documents will be characterised by a $P(U|k) = 1$ (or ≈ 1 for $uRBP_2$), thus reducing uRBP to RBP. The fact that in general $uRBP_2$ is correlated with RBP more than $uRBP_1$ is to RBP highlights the effect of smoothing obtained by the arctan function; specifically, the increase of readability scores for which $P(U|k)$ is not zero

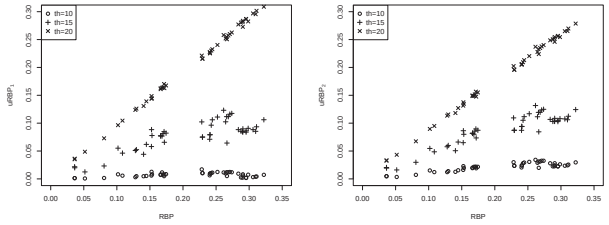

Figure 2: RBP vs. uRBP of CLEF eHealth 2013 systems (left: $uRBP_1$; right: $uRBP_2$) for varying values of threshold on the readability scores ($th = 10, 15, 20$).

	$th = 10$	$th = 15$	$th = 20$
RBP vs.	$\tau = .1277$	$\tau = .5603$	$\tau = .9574$
$uRBP_1$	$\tau_{AP} = -.0255$	$\tau_{AP} = .2746$	$\tau_{AP} = .9261$
RBP vs.	$\tau = .5887$	$\tau = .6791$	$\tau = .9574$
$uRBP_2$	$\tau_{AP} = .2877$	$\tau_{AP} = .4102$	$\tau_{AP} = .9407$

Table 1: Correlation coefficients (τ and τ_{AP}) between system rankings obtained with RBP and $uRBP_1$ or $uRBP_2$ for different values of the readability threshold.

beyond th narrows the scope for ranking differences between systems effectiveness. These observations are confirmed in Figure 2, where only few changes in the rank of systems are shown for $th = 20$ ($\times$ in Figure 2), while more changes are found for $th = 10$ ($\circ$) and $th = 15$ ($+$). Note that the small differences in the absolute values of effectiveness recorded by uRBP with $th = 10$ should not be interpreted as a lack of discriminative power. When $th = 10$ only 1.4% of the documents in the CLEF eHealth 2013 qrels are relevant and readable, thus contributing to uRBP.

Figure 2 demonstrates the importance of considering understandability along with topicality in the evaluation of systems for the considered task. The system ranked highest according to RBP (`MEDINFO.1.3.noadd`) is second to a number of systems according to uRBP if user understandability of up to FOG level 15 is wanted. Specifically, the highest $uRBP_1$ for $th = 10$ is achieved by `UTHealth_CCB.1.3.noadd`, which is ranked 28th according to RBP, and for $th = 15$ by `teamAEHRC.6.3`, which is ranked 19th according to RBP and achieves the highest $uRBP_2$ for $th = 10, 15$.

4. LIMITATIONS AND CONCLUSIONS

In this paper, we have investigated how understandability can be integrated in the gain-discount framework for evaluating IR systems. The approach studied here is general and can be adopted to other factors of relevance, such as reliability. Information reliability plays an important role in consumer health advice search; its integration will be studied in future work.

In the proposed approach, the relevance (P(R|k)) was modelled as the joint probability $P(T, U|k)$. This joint probability was assumed to be independent and the two events to be compositional, thus allowing to derive $P(T, U|k) = P(T|k)P(U|k)$ and to treat topicality and understandability separately. This is a strong assumption and it is not necessarily true; alternatives are under investigation, e.g. [2].

The approach was demonstrated by deriving understandability-based variants of RBP; other measures can also be extended, e.g., nDCG and ERR. Note, however, that nDCG-style versions would require normalising the gain function by the ideal gain, which in turns requires finding the optimal ranking based on two criteria, relevance score and understandability, instead of one as in the standard nDCG.

Xu and Chen [12] have noted that factors of relevance influence relevance assessments in different proportions, e.g., in their study, topicality was found to be more influential than understandability. The specific uRBP measures studied here did not consider this aspect; however weighting of different factors could be accomplished through a different $f(.)$ function for converting $P(T, U|k)$ into gain values.

In this paper, we have used readability as a proxy for understandability, but this is only one aspect that influences understandability [12]; future work may explore other factors, e.g., users' prior knowledge, as well as the presence of images that further explain the textual information. Furthermore, readability was estimated using general, surface level readability measures. Previous work has shown that these measures are often not suitable to evaluate the readability of health information. For example, Yan et al. [13] claim that people experience the highest readability difficulties at word level rather than at sentence level; they further propose a new metric based on concept-based readability, specifically instantiated in the health domain. A number of alternative approaches that measure text readability beyond the surface characteristics of text have been proposed. Future work will investigate their use to estimate P(U|k), along with actual readability assessments collected from users.

5. REFERENCES

[1] O. H. Ahmed, S. J. Sullivan, A. G. Schneiders, and P. R. McCrory. Concussion information online: evaluation of information quality, content and readability of concussion-related websites. *British journal of sports medicine*, 46(9):675–683, 2012.

[2] P. D. Bruza, G. Zuccon, and L. Sitbon. Modelling the information seeking user by the decision they make. In *MUBE 2013*, pages 5–6. ACM, 2013.

[3] B. Carterette. System effectiveness, user models, and user utility: a conceptual framework for investigation. In *SIGIR'11*, pages 903–912, 2011.

[4] C. L. Clarke, N. Craswell, and I. Soboroff. Overview of the TREC 2009 web track. In *TREC'09*, 2009.

[5] S. Fox and M. Duggan. Health online 2013. Tech. Rep., Pew Research Center's Internet & American Life Project, 2013.

[6] L. Goeuriot, G. Jones, L. Kelly, J. Leveling, A. Hanbury, H. Müller, S. Salanterä, H. Suominen, and G. Zuccon. Share/clef ehealth evaluation lab 2013, task 3: Information retrieval to address patients' questions when reading clinical reports. In *CLEF*, 2013.

[7] D. R. McCallum and J. L. Peterson. Computer-based readability indexes. In *ACM'82 Conf.*, pages 44–48, 1982.

[8] A. Moffat and J. Zobel. Rank-biased precision for measurement of retrieval effectiveness. *TOIS*, 27(1):2, 2008.

[9] M. D. Smucker and C. L. Clarke. Time-based calibration of effectiveness measures. In *SIGIR'12*, pages 95–104, 2012.

[10] T. M. Walsh and T. A. Volsko. Readability assessment of internet-based consumer health information. *Respiratory care*, 53(10):1310–1315, 2008.

[11] R. C. Wiener and R. Wiener-Pla. Literacy, pregnancy and potential oral health changes: The internet and readability levels. *Maternal and child health journal*, pages 1–6, 2013.

[12] Y. C. Xu and Z. Chen. Relevance judgment: What do information users consider beyond topicality? *JASIST*, 57(7):961–973, 2006.

[13] X. Yan, D. Song, and X. Li. Concept-based document readability in domain specific information retrieval. In *CIKM'06*, pages 540–549, 2006.

[14] E. Yilmaz, J. A. Aslam, and S. Robertson. A new rank correlation coefficient for information retrieval. In *SIGIR'08*, pages 587–594, 2008.